843.43861X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s):	T. OKAMOTO et aL
Serial No.:	Not Yet Assigned
Filed:	On Even Date Herewith
Title:	DISK ARRAY APPARATUS AND CONTROL METHOD FOR THE SAME

LETTER CLAIMING RIGHT OF PRIORITY

Mail Stop: Patent Applications
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

May 20, 2004

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby claim(s) the right of priority based on **Japanese** Patent Application No.**2004-063313**, filed March 8, 2004.

A certified copy of said Japanese Application is attached.

Respectfully submitted,
ANTONELLI, TERRY, STOUT & KRAUS, LLP

Carl I. Brundidge
Registration Number 29,621

CIB/dks
Attachment
(703) 312-6600

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００４年　３月　８日

出願番号
Application Number: 特願２００４－０６３３１３

[ST. 10/C]: [ＪＰ２００４－０６３３１３]

出願人
Applicant(s): 株式会社日立製作所

２００４年　４月　７日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００４－３０２８５６７

【書類名】　特許願
【整理番号】　K04001481
【提出日】　平成16年 3月 8日
【あて先】　特許庁長官殿
【国際特許分類】　G06F　3/06
【発明者】
　【住所又は居所】　神奈川県小田原市中里３２２番２号　株式会社日立製作所　ＲＡＩＤシステム事業部内
　【氏名】　岡本　岳樹
【発明者】
　【住所又は居所】　神奈川県小田原市中里３２２番２号　株式会社日立製作所　ＲＡＩＤシステム事業部内
　【氏名】　福岡　幹夫
【特許出願人】
　【識別番号】　000005108
　【氏名又は名称】　株式会社日立製作所
【代理人】
　【識別番号】　100080001
　【弁理士】
　【氏名又は名称】　筒井　大和
　【電話番号】　03-3366-0787
【手数料の表示】
　【予納台帳番号】　006909
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　特許請求の範囲　1
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1

【書類名】特許請求の範囲
【請求項１】
記憶領域に対するデータの書き込みまたは読み出し処理においてエラーが生じた場合、データの書き込みまたは読み出しが正常に終了したことを通知した後に、データの書き込みまたは読み出しを再度繰り返し、データの書き込みまたは読み出しを実行する複数の記憶デバイスと、
前記複数の記憶デバイスに対する書き込みデータまたは読み出し要求が格納される格納領域を有して、前記複数の記憶デバイスに対するデータの書き込みまたは読み出しを制御し、前記複数の記憶デバイスからデータの書き込みまたは読み出しが正常に終了したことを通知される記憶デバイス制御部と、
自ディスクアレイ装置の外部のネットワークから書き込み要求または読み出し要求を受けるチャネル制御部と、
前記チャネル制御部、前記記憶デバイス制御部によって通信される制御情報が格納される共有メモリと、
前記チャネル制御部と前記記憶デバイス制御部との間で通信されるデータが一時的に保存されるキャッシュメモリと、
前記チャネル制御部、前記記憶デバイス制御部、前記共有メモリおよび前記キャッシュメモリに接続される接続部とを有し、
前記記憶デバイス制御部は、
データの書き込みまたは読み出しに利用され、冗長性を有してデータを格納できる論理的な記憶領域を、前記複数の記憶デバイスの記憶領域を用いて生成し、
前記論理的な記憶領域を形成する複数の記憶デバイスに関して、前記複数の記憶デバイスに対する書き込みデータまたは読み出し要求が格納される格納領域を監視し、
前記論理的な記憶領域を形成する複数の記憶デバイスのうちデータの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定して、前記特定された記憶デバイスを閉塞させることを特徴とするディスクアレイ装置。
【請求項２】
請求項１記載のディスクアレイ装置において、
前記複数の記憶デバイスは、冗長性を有し、
前記共有メモリは、前記キャッシュメモリ上の前記記憶デバイスに書き込むべき未反映データ量を保持する領域を有し、
前記記憶デバイス制御部は、前記データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定する際に、冗長性を有する前記複数の記憶デバイスの未反映データ量を比較し、未反映データ量の多い記憶デバイスを閉塞させる記憶デバイスとして特定することを特徴とするディスクアレイ装置。
【請求項３】
請求項１記載のディスクアレイ装置において、
前記複数の記憶デバイスは、冗長性を有し、
前記共有メモリは、前記複数の記憶デバイスのそれぞれの平均応答時間を保持する領域を有し、
前記記憶デバイス制御部は、前記データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定する際に、冗長性を有する前記複数の記憶デバイスの平均応答時間を比較し、平均応答時間の大きい記憶デバイスを閉塞させる記憶デバイスとして特定することを特徴とするディスクアレイ装置。
【請求項４】
請求項１記載のディスクアレイ装置において、
前記複数の記憶デバイスは、冗長性を有し、
前記記憶デバイス制御部は、前記複数の記憶デバイスのそれぞれに対するキュー数を保持する領域を有し、前記データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定する際に、冗長性を有する前記複数の記憶デバイスのキュー数を比較し、キュ

一数が多い記憶デバイスを閉塞させる記憶デバイスとして特定することを特徴とするディスクアレイ装置。

【請求項５】

請求項１記載のディスクアレイ装置において、

前記複数の記憶デバイスは、冗長性を有し、

前記記憶デバイス制御部は、前記データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定して閉塞させた後、冗長性を有する他の記憶デバイスのデータを用いて予備の記憶デバイスにデータを復元することを特徴とするディスクアレイ装置。

【請求項６】

請求項１記載のディスクアレイ装置において、

前記複数の記憶デバイスは、冗長性を有し、

前記接続部に接続される管理端末をさらに有し、

前記管理端末は、冗長性を有する前記複数の記憶デバイスのうち前記データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定する際の条件を設定することを特徴とするディスクアレイ装置。

【請求項７】

請求項６記載のディスクアレイ装置において、

前記条件は、前記データの書き込みまたは読み出しの繰り返し回数であることを特徴とするディスクアレイ装置。

【請求項８】

請求項６記載のディスクアレイ装置において、

前記条件は、前記キャッシュメモリ上の前記記憶デバイスに書き込むべき未反映データ量の差分倍数であることを特徴とするディスクアレイ装置。

【請求項９】

請求項６記載のディスクアレイ装置において、

前記条件は、前記複数の記憶デバイスのそれぞれの平均応答時間の差分倍数であることを特徴とするディスクアレイ装置。

【請求項１０】

請求項６記載のディスクアレイ装置において、

前記条件は、前記複数の記憶デバイスのそれぞれに対するキュー数の差分倍数であることを特徴とするディスクアレイ装置。

【請求項１１】

記憶領域に対するデータの書き込みまたは読み出し処理においてエラーが生じた場合、データの書き込みまたは読み出しが正常に終了したことを通知した後に、データの書き込みまたは読み出しを再度繰り返し、データの書き込みまたは読み出しを実行する複数の記憶デバイスと、

前記複数の記憶デバイスに対する書き込みデータまたは読み出し要求が格納される格納領域を有して、前記複数の記憶デバイスに対するデータの書き込みまたは読み出しを制御し、前記複数の記憶デバイスからデータの書き込みまたは読み出しが正常に終了したことを通知される記憶デバイス制御部と、

自ディスクアレイ装置の外部のネットワークから書き込み要求または読み出し要求を受けるチャネル制御部と、

前記チャネル制御部、前記記憶デバイス制御部によって通信される制御情報が格納される共有メモリと、

前記チャネル制御部と前記記憶デバイス制御部との間で通信されるデータが一時的に保存されるキャッシュメモリと、

前記チャネル制御部、前記記憶デバイス制御部、前記共有メモリおよび前記キャッシュメモリに接続される接続部とを有するディスクアレイ装置の制御方法であって、

前記記憶デバイス制御部は、

データの書き込みまたは読み出しに利用され、冗長性を有してデータを格納できる論

理的な記憶領域を、前記複数の記憶デバイスの記憶領域を用いて生成し、
　前記論理的な記憶領域を形成する複数の記憶デバイスに関して、前記複数の記憶デバイスに対する書き込みデータまたは読み出し要求が格納される格納領域を監視し、
　前記論理的な記憶領域を形成する複数の記憶デバイスのうちデータの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定して、前記特定された記憶デバイスを閉塞させることを特徴とするディスクアレイ装置の制御方法。
【請求項１２】
　請求項１１記載のディスクアレイ装置の制御方法において、
　前記複数の記憶デバイスは、冗長性を有し、
　前記共有メモリは、前記キャッシュメモリ上の前記記憶デバイスに書き込むべき未反映データ量を保持する領域を有し、
　前記記憶デバイス制御部は、前記データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定する際に、冗長性を有する前記複数の記憶デバイスの未反映データ量を比較し、未反映データ量の多い記憶デバイスを閉塞させる記憶デバイスとして特定することを特徴とするディスクアレイ装置の制御方法。
【請求項１３】
　請求項１１記載のディスクアレイ装置の制御方法において、
　前記複数の記憶デバイスは、冗長性を有し、
　前記共有メモリは、前記複数の記憶デバイスのそれぞれの平均応答時間を保持する領域を有し、
　前記記憶デバイス制御部は、前記データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定する際に、冗長性を有する前記複数の記憶デバイスの平均応答時間を比較し、平均応答時間の大きい記憶デバイスを閉塞させる記憶デバイスとして特定することを特徴とするディスクアレイ装置の制御方法。
【請求項１４】
　請求項１１記載のディスクアレイ装置の制御方法において、
　前記複数の記憶デバイスは、冗長性を有し、
　前記記憶デバイス制御部は、前記複数の記憶デバイスのそれぞれに対するキュー数を保持する領域を有し、前記データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定する際に、冗長性を有する前記複数の記憶デバイスのキュー数を比較し、キュー数が多い記憶デバイスを閉塞させる記憶デバイスとして特定することを特徴とするディスクアレイ装置の制御方法。
【請求項１５】
　請求項１１記載のディスクアレイ装置の制御方法において、
　前記複数の記憶デバイスは、冗長性を有し、
　前記記憶デバイス制御部は、前記データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定して閉塞させた後、冗長性を有する他の記憶デバイスのデータを用いて予備の記憶デバイスにデータを復元することを特徴とするディスクアレイ装置の制御方法。
【請求項１６】
　請求項１１記載のディスクアレイ装置の制御方法において、
　前記複数の記憶デバイスは、冗長性を有し、
　前記接続部に接続される管理端末をさらに有し、
　前記管理端末は、冗長性を有する前記複数の記憶デバイスのうち前記データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定する際の条件を設定することを特徴とするディスクアレイ装置の制御方法。
【請求項１７】
　請求項１６記載のディスクアレイ装置の制御方法において、
　前記条件は、前記データの書き込みまたは読み出しの繰り返し回数であることを特徴とするディスクアレイ装置の制御方法。

【請求項１８】
請求項１６記載のディスクアレイ装置の制御方法において、
前記条件は、前記キャッシュメモリ上の前記記憶デバイスに書き込むべき未反映データ量の差分倍数であることを特徴とするディスクアレイ装置の制御方法。
【請求項１９】
請求項１６記載のディスクアレイ装置の制御方法において、
前記条件は、前記複数の記憶デバイスのそれぞれの平均応答時間の差分倍数であることを特徴とするディスクアレイ装置の制御方法。
【請求項２０】
請求項１６記載のディスクアレイ装置の制御方法において、
前記条件は、前記複数の記憶デバイスのそれぞれに対するキュー数の差分倍数であることを特徴とするディスクアレイ装置の制御方法。

【書類名】明細書
【発明の名称】ディスクアレイ装置およびその制御方法
【技術分野】
【０００１】
本発明は、ディスクアレイ装置およびその制御技術に関し、特に、性能劣化しているディスクドライブの検出に適用して有効な技術に関する。
【背景技術】
【０００２】
本発明者が検討したところによれば、従来のディスクアレイ装置およびその制御技術に関しては、以下のような技術が考えられる。
【０００３】
たとえば、従来のディスクアレイ装置およびその制御技術においては、ディスクドライブに対するデータの書き込みまたは読み出し処理においてエラーが生じた場合、データの書き込みまたは読み出しを繰り返して行う処理、いわゆるリトライが行われる。このリトライにおいては、発生したエラー要因を取得し、この取得したエラー要因に対応するリトライ回数を設定し、この回数以内で救済に成功した場合にはリカバードとしてホストに報告される（特許文献１参照）。
【０００４】
ところで、ディスクアレイ装置のディスクドライブは、年々大容量化に伴い記録密度も高密度化の一途を辿っている。そのため、記録密度当たりのエラーレートは、従来のディスクドライブと同等であっても、ディスクドライブ当たりでのエラー件数は記録密度に比例することになり、従来と同じようにリトライにより救済できた事例を毎々ホストに報告してしまうと、エラー報告件数が多大となり、すぐに故障ディスクドライブと判断されやすくなる。
【０００５】
その打開策の一例として、たとえばリトライ回数が数回であった場合には、ホストにリカバード（救済成功）ではなく、正常終了を報告するような方式、いわゆる隠しリトライを採用することがある。この隠しリトライにおいて、ディスクドライブを使用するホストは、通常、ディスクドライブの応答時間を監視しているが、突発的な一時障害を考慮し、監視時間は正常な応答時間の１０倍～１００倍程度に設定していることが常である。そのため、慢性的に救済処理を行い、応答時間が常に正常時の１．５～２倍程度に遅延しているディスクドライブを遅延と検出することはない。
【特許文献１】特開２００３－１４１８２４号公報
【発明の開示】
【発明が解決しようとする課題】
【０００６】
ところで、前記のような本発明者が検討した従来のディスクアレイ装置およびその制御技術に関して、本発明者が検討した結果、以下のようなことが明らかとなった。
【０００７】
たとえば、従来のディスクアレイ装置およびその制御技術においては、前述の通り、ディスクドライブはホストには見えない隠しリトライや、またはより複雑になったヘッド制御シーケンスのμ不良等により、ホストに正常終了と報告していても、その報告時間は正常時の数倍になることがあり、それが慢性的に発生すれば、慢性的な性能劣化ドライブとなる。このように、ホスト側が単純に正常時の報告時間の数倍を目処に監視時間を設定してしまうと、突発的な一時障害も検出してしまうため、これを区別するためには統計的に判断する必要がある。
【０００８】
そこで、本発明の目的は、ディスクドライブ自身に性能劣化を検出する機能がなくても、性能劣化しているディスクドライブを検出することができ、さらに性能劣化レベルを可変できることにより、顧客要求に応じたディスクアレイ装置の構築を実現することができ

る技術を提供することにある。

【０００９】

本発明の前記ならびにその他の目的と新規な特徴は、本明細書の記述および添付図面から明らかになるであろう。

【課題を解決するための手段】

【００１０】

本願において開示される発明のうち、代表的なものの概要を簡単に説明すれば、次のとおりである。

【００１１】

本発明は、記憶領域に対するデータの書き込みまたは読み出し処理においてエラーが生じた場合、データの書き込みまたは読み出しが正常に終了したことを通知した後に、データの書き込みまたは読み出しを再度繰り返し、データの書き込みまたは読み出しを実行する複数の記憶デバイスと、複数の記憶デバイスに対する書き込みデータまたは読み出し要求が格納される格納領域を有して、複数の記憶デバイスに対するデータの書き込みまたは読み出しを制御し、複数の記憶デバイスからデータの書き込みまたは読み出しが正常に終了したことを通知される記憶デバイス制御部と、自ディスクアレイ装置の外部のネットワークから書き込み要求または読み出し要求を受けるチャネル制御部と、チャネル制御部、記憶デバイス制御部によって通信される制御情報が格納される共有メモリと、チャネル制御部と記憶デバイス制御部との間で通信されるデータが一時的に保存されるキャッシュメモリと、チャネル制御部、記憶デバイス制御部、共有メモリおよびキャッシュメモリに接続される接続部とを有するディスクアレイ装置およびその制御方法に適用され、以下のような特徴を有するものである。

【００１２】

すなわち、本発明のディスクアレイ装置において、記憶デバイス制御部は、データの書き込みまたは読み出しに利用され、冗長性を有してデータを格納できる論理的な記憶領域を、複数の記憶デバイスの記憶領域を用いて生成し、論理的な記憶領域を形成する複数の記憶デバイスに関して、複数の記憶デバイスに対する書き込みデータまたは読み出し要求が格納される格納領域を監視し、論理的な記憶領域を形成する複数の記憶デバイスのうちデータの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定して、特定された記憶デバイスを閉塞させるものである。

【００１３】

具体的には、データの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定する際に、冗長性を有する複数の記憶デバイスの未反映データ量を比較し、未反映データ量の多い記憶デバイスを閉塞させる記憶デバイスとして特定したり、または冗長性を有する複数の記憶デバイスの平均応答時間を比較し、平均応答時間の大きい記憶デバイスを閉塞させる記憶デバイスとして特定したり、または冗長性を有する複数の記憶デバイスのキュー数を比較し、キュー数が多い記憶デバイスを閉塞させる記憶デバイスとして特定するものである。

【００１４】

さらに、接続部に接続される管理端末を有し、冗長性を有する複数の記憶デバイスのうちデータの書き込みまたは読み出しの繰り返し回数が多い記憶デバイスを特定する際の条件、たとえばデータの書き込みまたは読み出しの繰り返し回数、キャッシュメモリ上の記憶デバイスに書き込むべき未反映データ量の差分倍数、複数の記憶デバイスのそれぞれの平均応答時間の差分倍数、複数の記憶デバイスのそれぞれに対するキュー数の差分倍数、などを設定できるようにしたものである。

【発明の効果】

【００１５】

本願において開示される発明のうち、代表的なものによって得られる効果を簡単に説明すれば以下のとおりである。

【００１６】

本発明によれば、ディスクドライブ自身に性能劣化を検出する機能がなくても、性能劣化しているディスクドライブを検出することが可能であり、さらに性能劣化レベルを可変できることにより、顧客要求に応じたディスクアレイ装置の構築を実現することが可能となる。

【発明を実施するための最良の形態】

【００１７】

以下、本発明の実施の形態を図面に基づいて詳細に説明する。なお、実施の形態を説明するための全図において、同一の機能を有する部材には原則として同一の符号を付し、その繰り返しの説明は省略する。

【００１８】

＜本発明の概念＞

本発明は、記憶領域に対するデータの書き込みまたは読み出し処理においてエラーが生じた場合、データの書き込みまたは読み出しが正常に終了したことを通知した後に、データの書き込みまたは読み出しを再度繰り返し、データの書き込みまたは読み出しを実行する複数のディスクドライブ（記憶デバイス）と、複数のディスクドライブに対する書き込みデータまたは読み出し要求が格納される格納領域を有して、複数のディスクドライブに対するデータの書き込みまたは読み出しを制御し、複数のディスクドライブからデータの書き込みまたは読み出しが正常に終了したことを通知されるディスクコントローラ（記憶デバイス制御部）と、自ディスクアレイ装置の外部のネットワークから書き込み要求または読み出し要求を受けるチャネルコントローラ（チャネル制御部）と、チャネルコントローラ、ディスクコントローラによって通信される制御情報が格納される共有メモリと、チャネルコントローラとディスクコントローラとの間で通信されるデータが一時的に保存されるキャッシュメモリと、チャネルコントローラ、ディスクコントローラ、共有メモリおよびキャッシュメモリに接続されるスイッチ（接続部）と、スイッチに接続されるサービスプロセッサ（管理端末）を有するディスクアレイ装置に適用される。

【００１９】

＜ディスクアレイ装置の全体構成例＞

図１により、本発明の一実施の形態に係るディスクアレイ装置の全体構成の一例を説明する。図１は本実施の形態に係るディスクアレイ装置の全体構成を示す構成図である。

【００２０】

本実施の形態に係るディスクアレイ装置１００は、ディスク制御装置２００とディスク駆動装置３００とを備える。ディスク制御装置２００は、たとえばホストの情報処理装置４００から受信したコマンドに従ってディスク駆動装置３００に対する制御を行う。たとえば、ホストの情報処理装置４００からデータ入出力要求を受信して、ディスク駆動装置３００が備えるディスクドライブ３１０に記憶されるデータの読み書きを行う。また、ディスク制御装置２００は、たとえば管理クライアントの情報処理装置５００からディスクアレイ装置１００を管理するための各種コマンドを受信して、ディスクアレイ装置１００の様々な設定を行う。

【００２１】

ホスト、管理クライアントの情報処理装置４００，５００は、ＣＰＵやメモリを備えたコンピュータなどの情報機器である。情報処理装置４００，５００が備えるＣＰＵによって各種プログラムが実行されることにより、様々な機能が実現される。情報処理装置４００，５００は、たとえばパーソナルコンピュータやワークステーションであることもあるし、メインフレームコンピュータであることもある。特に、ホストの情報処理装置４００は、たとえば銀行の自動預金預け払いシステムや航空機の座席予約システムなどにおける中枢コンピュータとして利用される。また、管理クライアントの情報処理装置５００は、ディスクアレイ装置１００を保守、管理するための管理コンピュータとして利用される。

【００２２】

ここで、ホスト、管理クライアントの情報処理装置４００，５００は、異なるユーザの情報処理装置とすることができる。たとえば、ホストの情報処理装置（１），（２）４０

０、管理クライアントの情報処理装置（６）５００は、ユーザＡの情報処理装置、ホストの情報処理装置（３）～（５）４００、管理クライアントの情報処理装置（７）５００は、ユーザＢの情報処理装置とすることができる。また、管理クライアントの情報処理装置（８）５００は、ディスクアレイ装置１００の全体を管理する管理者の情報処理装置とすることができる。ここで、ユーザとは、たとえば企業とすることができ、または企業内における部署などの部門とすることもでき、あるいは個人とすることもできる。

【００２３】

図１において、ホストの情報処理装置４００は、ＳＡＮ（Ｓｔｏｒａｇｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）６００を介してディスク制御装置２００と通信可能に接続されている。ＳＡＮ６００は、ディスク駆動装置３００が提供する記憶資源におけるデータの管理単位であるブロックを単位としてホストの情報処理装置４００との間でデータの授受を行うためのネットワークである。ＳＡＮ６００を介して行われるホストの情報処理装置４００とディスク制御装置２００との間の通信は、たとえばファイバチャネルプロトコルに従って行われるようにすることができる。

【００２４】

もちろん、ホストの情報処理装置４００とディスク制御装置２００との間は、ＳＡＮ６００を介して接続されている必要はなく、たとえば、ＬＡＮ（Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）を介して接続されているようにすることもできるし、ネットワークを介さずに直接に接続されているようにすることもできる。ＬＡＮを介して接続される場合には、たとえばＴＣＰ／ＩＰ（Ｔｒａｎｓｍｉｓｓｏｎ　Ｃｏｎｔｒｏｌ　Ｐｒｏｔｏｃｏｌ／Ｉｎｔｅｒｎｅｔ　Ｐｒｏｔｏｃｏｌ）プロトコルに従って通信を行うようにすることができる。また、ネットワークを介さずに直接に接続される場合には、たとえばＦＩＣＯＮ（Ｆｉｂｒｅ　Ｃｏｎｎｅｃｔｉｏｎ）（登録商標）やＥＳＣＯＮ（Ｅｎｔｅｒｐｒｉｓｅ　Ｓｙｓｔｅｍ　Ｃｏｎｎｅｃｔｉｏｎ）（登録商標）、ＡＣＯＮＡＲＣ（Ａｄｖａｎｃｅｄ　Ｃｏｎｎｅｃｔｉｏｎ　Ａｒｃｈｉｔｅｃｔｕｒｅ）（登録商標）、ＦＩＢＡＲＣ（Ｆｉｂｒｅ　Ｃｏｎｎｅｃｔｉｏｎ　Ａｒｃｈｉｔｅｃｔｕｒｅ）（登録商標）などの通信プロトコルに従って通信を行うようにすることもできる。

【００２５】

また、管理クライアントの情報処理装置５００は、ＬＡＮ７００を介してディスク制御装置２００と接続されている。ＬＡＮ７００は、インターネットとすることもできるし、専用のネットワークとすることもできる。ＬＡＮ７００を介して行われる管理クライアントの情報処理装置５００とディスク制御装置２００との間の通信は、たとえばＴＣＰ／ＩＰプロトコルに従って行われるようにすることができる。

【００２６】

＜＜ディスク駆動装置＞＞

ディスク駆動装置３００は、多数のディスクドライブ３１０を備えている。これにより、ホストの情報処理装置４００、管理クライアントの情報処理装置５００に対して大容量の記憶領域を提供することができる。ディスクドライブ３１０は、ハードディスクドライブなどのデータ記憶媒体、あるいはＲＡＩＤ（Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙｓ　ｏｆ　Ｉｎｅｘｐｅｎｓｉｖｅ　Ｄｉｓｋｓ）を構成する複数のハードディスクドライブにより構成されてなるようにすることができる。また、ディスクドライブ３１０により提供される物理的な記憶領域である物理ボリュームには、論理的な記録領域である論理ボリュームを設定することができる。

【００２７】

ディスク制御装置２００とディスク駆動装置３００との間は、図１のように直接に接続される形態とすることもできるし、ネットワークを介して接続されるようにすることもできる。さらに、ディスク駆動装置３００は、ディスク制御装置２００と一体として構成されることもできる。

【００２８】

＜＜ディスク制御装置＞＞

ディスク制御装置２００は、チャネルコントローラ２１０、共有メモリ２２０、キャッシュメモリ２３０、ディスクコントローラ２４０、サービスプロセッサ２５０、スイッチ２６０を備える。ディスク制御装置２００は、チャネルコントローラ２１０によりＳＡＮ６００を介してホストの情報処理装置４００との間の通信を行う。

【００２９】

チャネルコントローラ２１０は、ホストの情報処理装置４００、管理クライアントの情報処理装置５００との間で通信を行うための通信インタフェースを備え、ホスト、管理クライアントの情報処理装置４００，５００との間でデータ入出力コマンドなどを授受する機能を備える。

【００３０】

各チャネルコントローラ２１０は、サービスプロセッサ２５０と共に内部ＬＡＮ２６１で接続されている。これにより、チャネルコントローラ２１０に実行させるマイクロプログラムなどをサービスプロセッサ２５０から送信し、インストールすることが可能となっている。

【００３１】

スイッチ２６０は、チャネルコントローラ２１０、共有メモリ２２０、キャッシュメモリ２３０、ディスクコントローラ２４０、サービスプロセッサ２５０を相互に接続する。チャネルコントローラ２１０、共有メモリ２２０、キャッシュメモリ２３０、ディスクコントローラ２４０、サービスプロセッサ２５０の間でのデータやコマンドの授受は、スイッチ２６０を介することにより行われる。スイッチ２６０は、たとえばクロスバスイッチで構成される。

【００３２】

共有メモリ２２０およびキャッシュメモリ２３０は、チャネルコントローラ２１０、ディスクコントローラ２４０により共有される記憶メモリである。共有メモリ２２０は、主に制御情報やコマンドなどを記憶するために利用されるのに対して、キャッシュメモリ２３０は、主にデータを記憶するために利用される。

【００３３】

たとえば、あるチャネルコントローラ２１０がホストの情報処理装置４００から受信したデータ入出力要求が書き込みコマンドであった場合には、当該チャネルコントローラ２１０は書き込みコマンドを共有メモリ２２０に書き込むと共に、ホストの情報処理装置４００から受信した書き込みデータをキャッシュメモリ２３０に書き込む。一方、ディスクコントローラ２４０は、共有メモリ２２０を監視しており、共有メモリ２２０に書き込みコマンドが書き込まれたことを検出すると、当該コマンドに従ってキャッシュメモリ２３０から書き込みデータを読み出してディスク駆動装置３００内のディスクドライブ３１０に書き込む。

【００３４】

また、あるチャネルコントローラ２１０がホストの情報処理装置４００から受信したデータ入出力要求が読み出しコマンドであった場合には、読み出し対象となるデータがキャッシュメモリ２３０に存在するかどうかを調べる。ここで、キャッシュメモリ２３０に存在すれば、チャネルコントローラ２１０はそのデータをホストの情報処理装置４００に送信する。一方、読み出し対象となるデータがキャッシュメモリ２３０に存在しない場合には、当該チャネルコントローラ２１０は読み出しコマンドを共有メモリ２２０に書き込むと共に、共有メモリ２２０を監視する。読み出しコマンドが共有メモリ２２０に書き込まれたことを検出したディスクコントローラ２４０は、ディスク駆動装置３００内のディスクドライブ３１０から読み出し対象となるデータを読み出して、これをキャッシュメモリ２３０に書き込むと共に、その旨を共有メモリ２２０に書き込む。そして、チャネルコントローラ２１０は、読み出し対象となるデータがキャッシュメモリ２３０に書き込まれたことを検出すると、そのデータをホストの情報処理装置４００に送信する。

【００３５】

このように、チャネルコントローラ２１０およびディスクコントローラ２４０の間では

、キャッシュメモリ２３０を介してデータの授受が行われ、キャッシュメモリ２３０には、ディスクドライブ３１０に記憶されるデータのうち、チャネルコントローラ２１０やディスクコントローラ２４０により読み書きされるデータが記憶される。

【００３６】

なお、チャネルコントローラ２１０からディスクコントローラ２４０に対するデータの書き込みや読み出しの指示を、共有メモリ２２０を介在させて間接的に行う構成の他、たとえばチャネルコントローラ２１０からディスクコントローラ２４０に対してデータの書き込みや読み出しの指示を共有メモリ２２０を介さずに直接に行う構成とすることもできる。また、チャネルコントローラ２１０にディスクコントローラ２４０の機能を持たせて、データ入出力制御部とすることもできる。

【００３７】

ディスクコントローラ２４０は、データを記憶する複数のディスクドライブ３１０と通信可能に接続され、ディスク駆動装置３００の制御を行う。たとえば、上述のように、チャネルコントローラ２１０がホストの情報処理装置４００から受信したデータ入出力要求に応じて、ディスクドライブ３１０に対してデータの読み書きを行う。

【００３８】

各ディスクコントローラ２４０は、サービスプロセッサ２５０と共に内部ＬＡＮ２６１で接続されており、相互に通信を行うことが可能である。これにより、ディスクコントローラ２４０に実行させるマイクロプログラムなどをサービスプロセッサ２５０から送信し、インストールすることが可能となっている。

【００３９】

本実施の形態においては、共有メモリ２２０およびキャッシュメモリ２３０がチャネルコントローラ２１０およびディスクコントローラ２４０に対して独立に設けられていることについて記載したが、これに限られるものではなく、共有メモリ２２０またはキャッシュメモリ２３０がチャネルコントローラ２１０およびディスクコントローラ２４０の各々に分散されて設けられることも好ましい。この場合、スイッチ２６０は、分散された共有メモリ２２０またはキャッシュメモリ２３０を有するチャネルコントローラ２１０およびディスクコントローラ２４０を相互に接続させることになる。

【００４０】

また、チャネルコントローラ２１０、ディスクコントローラ２４０、スイッチ２６０、共有メモリ２２０、キャッシュメモリ２３０の少なくともいずれかが一体として構成されているようにすることもできる。

【００４１】

サービスプロセッサ２５０は、ディスクアレイ装置１００を保守・管理するためのコンピュータである。オペレータは、サービスプロセッサ２５０を操作することにより、たとえばディスク駆動装置３００内のディスクドライブ３１０の構成の設定や、ホストの情報処理装置４００、管理クライアントの情報処理装置５００とチャネルコントローラ２１０との間の通信路であるパスの設定、論理ボリュームの設定、チャネルコントローラ２１０やディスクコントローラ２４０において実行されるマイクロプログラムのインストールなどを行うことができる。ここで、ディスク駆動装置３００内のディスクドライブ３１０の構成の設定としては、たとえばディスクドライブ３１０の増設や減設、ＲＡＩＤ構成の変更（ＲＡＩＤ１からＲＡＩＤ５への変更など）などとすることができる。

【００４２】

さらに、サービスプロセッサ２５０からは、ディスクアレイ装置１００の動作状態の確認や故障部位の特定、チャネルコントローラ２１０で実行されるオペレーティングシステムのインストールなどの作業を行うことができる。これらの設定や制御は、サービスプロセッサ２５０が備えるユーザインタフェース、あるいはサービスプロセッサ２５０で動作するＷｅｂサーバにより提供されるＷｅｂページを表示する管理クライアントの情報処理装置５００のユーザインタフェースからオペレータなどにより行うようにすることができる。オペレータなどは、サービスプロセッサ２５０を操作して障害監視する対象や内容の

設定、障害通知先の設定などを行うこともできる。

【００４３】

サービスプロセッサ２５０は、ディスク制御装置２００に内蔵されている形態とすることもできるし、外付けされている形態とすることもできる。また、サービスプロセッサ２５０は、ディスク制御装置２００およびディスク駆動装置３００の保守・管理を専用に行うコンピュータとすることもできるし、汎用のコンピュータに保守・管理機能を持たせたものとすることもできる。

【００４４】

＜ディスクアレイ装置の外観構成例＞

図２および図３により、本発明の一実施の形態に係るディスクアレイ装置の外観構成の一例を説明する。それぞれ、図２は本実施の形態に係るディスクアレイ装置の外観構成を示す図、図３はディスク制御装置の外観構成を示す図、である。

【００４５】

図２に示すように、本実施の形態に係るディスクアレイ装置１００は、ディスク制御装置２００およびディスク駆動装置３００がそれぞれの筐体に納められた形態をしている。図２に示す例では、ディスク制御装置２００の筐体の両側にディスク駆動装置３００の筐体が配置されている。

【００４６】

ディスク制御装置２００は、正面中央部にサービスプロセッサ２５０が備えられている。サービスプロセッサ２５０はカバーで覆われており、図３に示すように、カバーを開けることにより、サービスプロセッサ２５０を使用することができる。なお、図３に示したサービスプロセッサ２５０は、いわゆるノート型パーソナルコンピュータの形態をしているが、どのような形態とすることも可能である。

【００４７】

サービスプロセッサ２５０の下部には、チャネルコントローラ２１０やディスクコントローラ２４０、キャッシュメモリ２３０、共有メモリ２２０、スイッチ２６０を装着するためのスロットが設けられている。チャネルコントローラ２１０やディスクコントローラ２４０、キャッシュメモリ２３０、共有メモリ２２０、スイッチ２６０は、回路基板を備えてボードとして構成されており、これらのボードが各スロットに装着される。各スロットには、これらのボードを装着するためのガイドレールが設けられている。ガイドレールに沿って各ボードをスロットに挿入することにより、チャネルコントローラ２１０やディスクコントローラ２４０、キャッシュメモリ２３０、共有メモリ２２０、スイッチ２６０をディスク制御装置２００に装着することができる。各スロットの奥手方向正面部には、各ボードをディスク制御装置２００と電気的に接続するためのコネクタが設けられている。

【００４８】

また、ディスク制御装置２００には、チャネルコントローラ２１０やディスクコントローラ２４０などから発生する熱を放出するためのファン２７０が設けられている。ファン２７０は、ディスク制御装置２００の上面部に設けられる他、スロットの上部にも設けられている。

【００４９】

＜ディスクアレイ装置の具体的構成および動作原理例＞

図４～図８により、本発明の一実施の形態に係るディスクアレイ装置の具体的構成および動作原理の一例を説明する。それぞれ、図４は本実施の形態に係るディスクアレイ装置の具体的構成を示す図、図５はホストコンピュータからのリード要求時のチャネルコントローラの動作を示すフローチャート、図６はチャネルコントローラからのリード要求時のディスクコントローラの動作を示すフローチャート、図７はホストコンピュータからのライト要求時のチャネルコントローラの動作を示すフローチャート、図８はチャネルコントローラからのライト要求時のディスクコントローラの動作を示すフローチャート、である。

【００５０】
本実施の形態に係るディスクアレイ装置１００は、具体的にはディスクサブシステムに適用され、図４に示すように、ホストコンピュータ（＝ホストの情報処理装置）４００に接続され、ホストインタフェースを具備したチャネルコントローラ２１０、共有メモリ２２０、キャッシュメモリ２３０、ＦＣ－ＡＬ（Ｆｉｂｒｅ　Ｃｈａｎｎｅｌ－Ａｒｂｉｔｒａｔｅｄ　Ｌｏｏｐ）インタフェースを例とするディスクドライブインタフェースを具備したディスクコントローラ２４０、サービスプロセッサ２５０と接続するためのサービスプロセッサインタフェース２５１、スイッチ２６０、ディスクドライブ３１０などから構成される。
【００５１】
ディスクドライブ３１０は、図４においてはディスクドライブ３１１～３１５から構成される例を示しており、ディスクドライブ（Ｄ１）３１１、ディスクドライブ（Ｄ２）３１２、ディスクドライブ（Ｄ３）３１３、およびディスクドライブ（Ｐ）３１４はＲＡＩＤ構成により冗長度を有しており、ディスクドライブ（Ｄ１）～（Ｄ３）はデータを格納するためのドライブとして備えられ、ディスクドライブ（Ｐ）はパリティ情報を格納するためのドライブとして備えられ、またディスクドライブ（Ｓ）は予備のドライブとして備えられている。
【００５２】
また、図４においては、サービスプロセッサ２５０が、サービスプロセッサインタフェース２５１を介してディスクサブシステムに外付けされている形態を示している。
【００５３】
本実施の形態のディスクサブシステムにおいては、特に、ディスクドライブ３１０（３１１～３１５）は、記憶領域に対するデータの書き込みまたは読み出し処理においてエラーが生じた場合、データの書き込みまたは読み出しが正常に終了したことを通知した後に、データの書き込みまたは読み出しを再度繰り返し、データの書き込みまたは読み出しを実行する機能を有している。
【００５４】
また、ディスクコントローラ２４０は、複数のディスクドライブ３１０に対する書き込みデータまたは読み出し要求が格納される格納領域を有して、複数のディスクドライブ３１０に対するデータの書き込みまたは読み出しを制御し、複数のディスクドライブ３１０からデータの書き込みまたは読み出しが正常に終了したことを通知されるとともに、特にデータの書き込みまたは読み出しに利用され、冗長性を有してデータを格納できる論理的な記憶領域を、複数のディスクドライブ３１０の記憶領域を用いて生成し、論理的な記憶領域を形成する複数のディスクドライブ３１０に関して、複数のディスクドライブ３１０に対する書き込みデータまたは読み出し要求が格納される格納領域を監視し、論理的な記憶領域を形成する複数のディスクドライブ３１０のうちデータの書き込みまたは読み出しの繰り返し回数が多いディスクドライブを特定して、特定されたディスクドライブを閉塞させる機能を有している。
【００５５】
さらに、チャネルコントローラ２１０は、自ディスクアレイ装置の外部のネットワークから書き込み要求または読み出し要求を受ける機能を有している。共有メモリ２２０は、チャネルコントローラ２１０、ディスクコントローラ２４０によって通信される制御情報が格納される領域を有している。キャッシュメモリ２３０は、チャネルコントローラ２１０とディスクコントローラ２４０との間で通信されるデータが一時的に保存される領域を有している。スイッチ２６０は、チャネルコントローラ２１０、ディスクコントローラ２４０、共有メモリ２２０およびキャッシュメモリ２３０に接続される。
【００５６】
また、サービスプロセッサ２５０は、冗長性を有する複数のディスクドライブ３１０のうちデータの書き込みまたは読み出しの繰り返し回数が多いディスクドライブを特定する際の条件を設定する機能を有している。

【００５７】
このように構成されるディスクサブシステムにおける動作は、以下の通りである。ホストコンピュータ４００からライトデータを受領したチャネルコントローラ２１０は、キャッシュメモリ２３０に退避すると共にディスクコントローラ２４０に対して、キャッシュメモリ２３０にあるライトデータをディスクドライブ３１１～３１４に書き込むように指示する。また、ホストコンピュータ４００からデータリード要求を受領したチャネルコントローラ２１０は、ディスクコントローラ２４０に対し、ディスクドライブ３１１～３１４よりデータを読み出し、キャッシュメモリ２３０に転送するように指示する。指示を受けたディスクコントローラ２４０はディスクドライブ３１１～３１４よりデータを読み出し、キャッシュメモリ２３０に転送した後、チャネルコントローラ２１０にデータ読み出し完了を報告する。報告を受けたチャネルコントローラ２１０はデータをキャッシュメモリ２３０よりホストコンピュータ４００に転送する。具体的には、以下の通りである。

【００５８】
具体的に、ホストコンピュータ４００からのリード要求時のチャネルコントローラ２１０の動作は、図５に示すように、Ｓ１００１でスイッチ２６０を介し、ディスクコントローラ２４０にリード要求を行う。次に、Ｓ１００２でディスクコントローラ２４０の応答監視を行い、応答がなかった場合（Ｎｏ）はＳ１００２へ、応答があった場合（Ｙｅｓ）はＳ１００３へ移行する。Ｓ１００３でキャッシュメモリ２３０にあるデータをスイッチ２６０を介して読み出し、ホストコンピュータ４００に転送し、転送が完了したら続くＳ１００４でホストコンピュータ４００に完了報告を行う。

【００５９】
チャネルコントローラ２１０からのリード要求時のディスクコントローラ２４０の動作は、図６に示すように、Ｓ１０１１でディスクドライブ３１１～３１５にリード要求を行う。次に、Ｓ１０１２で転送開始を監視し、転送開始要求があった場合（Ｙｅｓ）、続くＳ１０１３でディスクドライブ３１１～３１５のデータを読み出し、スイッチ２６０を介してキャッシュメモリ２３０に転送する。続く、Ｓ１０１４で転送完了を監視し、転送が完了したら（Ｙｅｓ）、Ｓ１０１５でチャネルコントローラ２１０にスイッチ２６０を介して完了報告を行う。

【００６０】
ホストコンピュータ４００からのライト要求時のチャネルコントローラ２１０の動作は、図７に示すように、Ｓ１０２１でスイッチ２６０を介し、ホストコンピュータ４００からのライトデータをキャッシュメモリ２３０に転送する。次に、Ｓ１０２２でディスクコントローラ２４０へスイッチ２６０を介してライト要求を行い、続くＳ１０２３でホストコンピュータ４００に完了報告を行う。

【００６１】
チャネルコントローラ２１０からのライト要求時のディスクコントローラ２４０の動作は、図８に示すように、Ｓ１０３１でキャッシュメモリ２３０よりスイッチ２６０を介してライトするデータを読み出し、冗長データを作成し、この作成した冗長データをスイッチ２６０を介してキャッシュメモリ２３０内の別の領域に書き込む。次に、Ｓ１０３２でディスクドライブ３１１～３１５にライト要求を行い、続くＳ１０３３で転送開始を監視し、転送開始要求があった場合（Ｙｅｓ）、続くＳ１０３４でキャッシュメモリ２３０のデータをディスクドライブ３１１～３１５に転送し、続くＳ１０３５で転送完了を監視し、転送が完了したら（Ｙｅｓ）、処理を終了する。

【００６２】
＜ディスクアレイ装置の入出力処理例＞
図９により、本発明の一実施の形態に係るディスクアレイ装置の入出力処理の一例を説明する。図９はディスクドライブに対する入出力処理を示す図である。

【００６３】
ディスクアレイ装置の入出力処理において、キャッシュメモリ２３０、共有メモリ２２０、ディスココントローラ２４０のローカルメモリ２４１にはそれぞれ、以下のような情

報が格納される。

【００６４】

キャッシュメモリ２３０には、ディスクドライブ３１１～３１３に対して書き込む未反映データ（Ｄ１）２３０１，（Ｄ２）２３０２，（Ｄ３）２３０３、ディスクドライブ３１４に書き込む冗長データ（Ｐ）２３０４、スペアのディスクドライブ３１５に書き込む再生データ（Ｓ）２３０５が格納される各領域が設けられている。さらに、キャッシュメモリ２３０には、ディスクドライブ３１１～３１５から読み出したリードデータ（Ｄ１’）２３１１，（Ｄ２’）２３１２，（Ｄ３’）２３１３，（Ｐ’）２３１４，（Ｓ’）２３１５が格納される各領域が設けられている。

【００６５】

共有メモリ２２０には、ディスクドライブ３１１～３１５に対する未反映データ量のドライブ情報（Ｄ１）２２０１，（Ｄ２）２２０２，（Ｄ３）２２０３，（Ｐ）２２０４，（Ｓ）２２０５や、ディスクドライブの性能劣化検出用のパラメータ２２０６の情報が格納される各領域が設けられている。

【００６６】

ディスココントローラ２４０のローカルメモリ２４１には、ディスクドライブ３１１～３１５に対応したキュー（Ｄ１）２４１１，（Ｄ２）２４１２，（Ｄ３）２４１３，（Ｐ）２４１４，（Ｓ）２４１５の情報が格納される各領域が設けられている。

【００６７】

図９において、ディスクドライブ３１１～３１３との入出力処理に関しては、チャネルコントローラ２１０よりディスクドライブ３１１～３１３に対して書き込み指示のあったデータは、キャッシュメモリ２３０内のディスクドライブ３１１～３１３に対応する未反映データ２３０１～２３０３に格納され、これらのデータを元にディスクドライブ３１４に書き込むべき冗長データ２３０４を作成し、未反映データ量は共有メモリ２２０にドライブ情報２２０１～２２０４として更新される。

【００６８】

ディスクコントローラ２４０は、ディスクドライブ３１１～３１４に対し、データの書き込みが完了したら、キャッシュメモリ２３０内のディスクドライブ３１１～３１４に対応する未反映データ２３０１～２３０４を破棄し、共有メモリ２２０内のドライブ情報２２０１～２２０４の未反映データ量を更新する。

【００６９】

たとえば一例として、ディスクドライブ（Ｄ２）３１２に対する書き込み指示があったとき、このディスクドライブ３１２のドライブ情報２２０２のドライブ状態が使用不可能であった場合は、冗長データ２３０４を作成した後、ディスクドライブ３１２のドライブ情報２２０２に対するライト要求を行わず、未反映データ２３０２およびドライブ情報２２０２の未反映データ量を削除する。

【００７０】

また一例として、ディスクドライブ（Ｄ１）３１１に対する読み出し指示があったとき、このディスクドライブ３１１のドライブ情報２２０１のドライブ状態が使用不可能であった場合は、冗長度を有するディスクドライブ３１２～３１４よりデータを読み出し、キャッシュメモリ２３０上のリードデータ２３１２～２３１４に格納し、これらのデータよりディスクドライブ３１１のリードデータ２３０１を再生する。

【００７１】

＜キューの構成例＞

図１０により、本発明の一実施の形態に係るディスクアレイ装置において、ディスクコントローラとディスクドライブで授受される入出力処理に関するキューの一例を説明する。図１０はキューの構成を示す図である。

【００７２】

ディスクコントローラ２４０とディスクドライブ３１１～３１５で授受され、各ディスクドライブ３１１～３１５に対応したキュー２４１１～２４１５は、ディスクコントロー

ラ２４０のローカルメモリ２４１上に格納される。それぞれは、個数を示すキュー数２４２１と、Ｎ１組の、コマンドの種別（ＲＥＡＤ、ＷＲＩＴＥなど）を示すコマンド種別２４２２、ディスクドライブのどの位置から入出力を開始するかの位置情報を示すＬＢＡ（Ｌｏｇｉｃａｌ　Ｂｌｏｃｋ　Ａｄｄｒｅｓｓ）の入出力開始位置２４２３、入出力のデータ転送量を示すデータ転送量２４２４、入出力を要求した時間を示す要求開始時間２４２５、当該キューが使用できるか否かを示す使用情報２４２６からなる個別キューより構成される。

【００７３】

たとえば、ディスクコントローラ２４０はディスクドライブ３１１～３１５に対し、ＦＣ－ＡＬ　Ｉ／Ｆを例とするインタフェースを介し、ＲＥＡＤまたはＷＲＩＴＥ要求を行った際、各ディスクドライブ毎のキュー２４１１～２４１５の個別キュー（コマンド種別、入出力開始位置、データ転送量、要求開始時間、使用情報）２４２２～２４２６を登録するとともに、キュー数２４２１をインクリメントする。また、ディスクドライブ３１１～３１５よりＲＥＡＤまたはＷＲＩＴＥ要求の応答があった時に、対応する各々のキュー２４１１～２４１５の個別キュー２４２２～２４２６を削除するとともに、キュー数２４２１をデクリメントする。

【００７４】

＜ドライブ情報の構成例＞

図１１～図１３により、本発明の一実施の形態に係るディスクアレイ装置において、共有メモリ上に格納されるドライブ情報の一例を説明する。それぞれ、図１１はドライブ情報の構成を示す図、図１２はディスクドライブが使用不可能時のリード要求の動作を示すフローチャート、図１３はディスクドライブが使用不可能時のライト要求の動作を示すフローチャート、である。

【００７５】

共有メモリ２２０上に格納されるドライブ情報２２０１～２２０５には、図１１に示すように、キャッシュメモリ２３０上にあって、ディスクドライブ３１１～３１５へのライト未反映データの量を示す未反映データ量２３２１と、応答時間テーブルの位置を示す応答時間ポインタ２３２２と、ディスクドライブ３１１～３１５に入出力を要求した要求開始時間２４２５から実際に応答があるまでの時間を示す応答時間（１）２３３１～（ｍ）２３３ｍと、応答時間２３３１～２３３ｍの平均を示す平均応答時間２３２３と、当該ディスクドライブが使用可能か否かを示すドライブ状態２３２４から構成される。

【００７６】

たとえば一例として、冗長構成を持つディスクドライブ３１１～３１４のうち、ディスクドライブ（Ｄ１）３１１が使用不可能時のリード要求の動作は、図１２に示すように、Ｓ１０４１で共有メモリ２２０上のドライブ情報２２０１の使用情報２４２６を参照し、ディスクドライブ３１１が使用不可能かどうかを判断する。使用可能な場合（Ｎｏ）、Ｓ１０４２でディスクドライブ（Ｄ１）３１１よりデータリードを行い、読み出したデータをキャッシュメモリ２３０上のリードデータ（Ｄ１’）２３１１の領域に格納する。Ｓ１０４１で使用不可能と判断した場合（Ｙｅｓ）は、Ｓ１０４３で冗長構成を持つ他のディスクドライブ（Ｄ２）３１２，（Ｄ３）３１３，（Ｐ）３１４よりデータリードを行い、キャッシュメモリ２３０上のリードデータ（Ｄ２’）２３１２，（Ｄ３’）２３１３，（Ｐ’）２３１４の領域にそれぞれ格納する。続くＳ１０４４で、これらのデータよりディスクドライブ（Ｄ１）３１１のデータを再生し、キャッシュメモリ２３０上のリードデータ（Ｄ１’）２３１１に格納する。

【００７７】

また一例として、冗長構成を持つディスクドライブ３１１～３１４のうち、ディスクドライブ（Ｄ１）３１１が使用不可能時のライト要求の動作は、図１３に示すように、Ｓ１０５１で共有メモリ２２０上のドライブ情報２２０１の使用情報２４２６を参照し、ディスクドライブ（Ｄ１）３１１が使用不可能かどうかを判断する。使用可能な場合（Ｎｏ）、Ｓ１０５２でキャッシュメモリ２３０上の未反映データ（Ｄ１）２３０１よりデータを

読み出してディスクドライブ（Ｄ１）３１１に書き込む。Ｓ１０５１で使用不可能と判断した場合（Ｙｅｓ）は、処理を終了する。

【００７８】

＜ディスクドライブの性能劣化検出例＞

図１４～図１６により、本発明の一実施の形態に係るディスクアレイ装置において、ディスクドライブの性能劣化検出の一例を説明する。それぞれ、図１４は共有メモリ上に格納されるディスクドライブの性能劣化検出用パラメータを示す図、図１５はディスクコントローラとディスクドライブの入出力シーケンスの動作を示すフローチャート、図１６（ａ）～（ｃ）は具体的な未反映データとデータ転送量の増減の関係を示す図、である。

【００７９】

共有メモリ２２０上に格納されるディスクドライブの性能劣化検出用パラメータ２２０６には、図１４に示すように、未反映データ量の差分倍数（ｎ１）２２１１、キュー数の差分倍数（ｎ２）２２１２、平均応答時間の差分倍数（ｎ３）２２１３などがある。

【００８０】

ディスクコントローラ２４０とディスクドライブ３１１～３１５の入出力シーケンスの動作は、図１５に示すように、Ｓ１０６１でディスクコントローラ２４０はディスクドライブ３１１～３１５よりデータ転送要求があるかどうかを判断する。データ転送要求がなかった場合（Ｎｏ）は、Ｓ１０６２でチャネルコントローラ２１０よりディスクドライブ３１１～３１５に対する入出力指示があったかどうかをチェックする。入出力指示がなかった場合（Ｎｏ）は、Ｓ１０６１に戻る。

【００８１】

Ｓ１０６２で入出力指示があった場合（Ｙｅｓ）は、Ｓ１０６３で入出力を発行するディスクドライブ３１１～３１５に対応する登録キュー数Ｍ１が登録可能キュー数Ｎ１未満かどうかをチェックし、未満でなければ（Ｎｏ）Ｓ１０６１に戻り、未満であれば（Ｙｅｓ）、Ｓ１０６４でキュー数Ｍ１をインクリメント（登録）し、１～Ｎ１までの間で使用情報２４２６が未使用のＰ１番目のキュー２４１１～２４１５にコマンド種別２４２２、入出力開始位置２４２３、データ転送量２４２４、現在の時間、すなわち要求開始時間２４２５を登録し、Ｐ１番目のキューの使用情報２４２６を使用中とし、Ｓ１０６５で、登録したキュー情報を元に対応するディスクドライブ３１１～３１５にキュー番号Ｐ１、コマンド種別２４２２、入出力開始位置２４２３、データ転送量２４２４の情報を送信し、入出力処理を要求する。

【００８２】

Ｓ１０６１でデータ転送要求があった場合（Ｙｅｓ）は、Ｓ１０６６で転送要求のあった情報に対応するキュー番号Ｐ１のキュー２４１１～２４１５の内容に従い、データ転送を実行する。たとえば、リード転送の場合は、ディスクドライブ３１１～３１５のデータをキャッシュメモリ２３０上の対応する位置の未反映データ２３０１～２３０５に転送し、ライト転送の場合は、キャッシュメモリ２３０に対応する位置の未反映データ２３０１～２３０５からディスクドライブ３１１～３１５に転送する。

【００８３】

次に、Ｓ１０６７でデータ転送が完了したかを確認し、未完了であれば（Ｎｏ）ステップ１０６６に戻り、完了していれば（Ｙｅｓ）、続くＳ１０６８でドライブ情報を更新する。すなわち、応答時間（＝［現在時間］－［キュー番号Ｐ１のキューに登録された要求時間］）を応答時間ポインタ（Ｘ）２３２２の示す応答時間２３３ｘに登録し、応答時間ポインタをインクリメントし（応答時間ポインタＸ＞ｍとなった場合は応答時間ポインタＸを０に戻す）、応答時間２３３１～２３３ｍより平均応答時間２３２３を算出する。また、完了したデータ転送がライト転送であった場合には、未反映データ量２３２１より転送完了したデータ転送量２４２４を除算する。次に、Ｓ１０６９で転送完了したキュー番号Ｐ１のキューの使用情報２４２６を未使用（削除）とし、登録キュー数Ｍ１をデクリメントする。次に、Ｓ１０７０でディスクドライブの応答遅延監視処理を実施し、Ｓ１０６１に戻る。

【００８４】
具体的に、未反映データとデータ転送量の増減の関係を、たとえば一例として、ディスクドライブ（Ｄ１）３１１を例に説明する。図１６において、（ａ）に示すように、キャッシュメモリ２３０上のディスクドライブ（Ｄ１）３１１の未反映データＤＤ０１，ＤＤ０２が存在したとすると、共有メモリ２２０上のドライブ情報（Ｄ１）２２０１の未反映データ量２３２１はＤＤ０１のサイズとＤＤ０２のサイズの和となる。次に、（ｂ）に示すように、新たなライト要求が来て、未反映データＤＤ０３がキャッシュメモリ２３０上に追加された場合、未反映データ量２３２１はＤＤ０１のサイズとＤＤ０２のサイズとＤＤ０３のサイズの和となる。次に、（ｃ）に示すように、新たなＤＤ０２の未反映データがディスクドライブに書き込み完了し、未反映データＤＤ０２がキャッシュメモリ２３０よりクリアされた場合は、未反映データ量２３２１はＤＤ０１のサイズとＤＤ０３のサイズの和となる。
【００８５】
ここでは、ディスクドライブ（Ｄ１）３１１について説明したが、ディスクドライブ（Ｄ２）３１２，（Ｄ３）３１３，（Ｐ）３１４，（Ｓ）３１５に関しても同様である。つまり、ディスクドライブの性能が劣化し、ディスクドライブの書き込み完了が遅延した場合、未反映データのクリアも遅れるため、未反映データ量が多くなることとなる。
【００８６】
＜ディスクドライブの応答遅延監視処理例＞
図１７～図２０により、本発明の一実施の形態に係るディスクアレイ装置において、ディスクドライブの応答遅延監視処理の一例を説明する。それぞれ、図１７は未反映データ量に着目した応答遅延監視処理を示すフローチャート、図１８は警告メッセージの表示画面を示す図、図１９はキュー数に着目した応答遅延監視処理を示すフローチャート、図２０は平均応答時間に着目した応答遅延監視処理を示すフローチャート、である。
【００８７】
本実施の形態のディスクサブシステムは、ＲＡＩＤ論理を用いた冗長度を有する構成になっており、たとえば前記図４では構成の３データドライブ＋１パリティドライブ（以下３Ｄ＋１Ｐと略す）の構成を示している。この構成では、通常、ホストコンピュータ４００の一定のサイズ以上のライトデータは均等に３分割され、ディスクドライブ（Ｄ１）３１１～（Ｄ３）３１３に書き込まれ、３分割されたデータより算出した冗長データはディスクドライブ（Ｐ１）３１４に書き込まれる。従って、１回のホストコンピュータ４００からのライト要求に対するディスクドライブ３１１～３１４に対して書き込みを行うべきデータ量は均一であり、それぞれのディスクドライブの性能が同等であれば、積算される未反映データ量も均等となる。もし、性能的に劣るディスクドライブが存在すれば、そのディスクドライブに対して、積算された未反映データ量は他のディスクドライブの未反映データ量よりも多くなることになる。
【００８８】
このことより、未反映データ量を考慮し、この未反映データ量に着目した応答遅延監視処理を図１７により説明する。図１７では、前記図４に示す構成（３Ｄ＋１Ｐ）でディスクドライブ（Ｄ２）３１２のデータ転送完了後のディスクドライブの遅延処理を例に説明する。
【００８９】
Ｓ１０７１で他のディスクドライブ３１１，３１３，３１４の共有メモリ２２０上のドライブ情報２２０１，２２０３，２２０４の未反映データ量２３２１の平均Ｑ１を求める。次に、Ｓ１０７２でディスクドライブ（Ｄ２）３１２の未反映データ量２３２１とＱ１（Ｑ１×未反映データ量差分倍数（ｎ１））を比較し、“未反映データ量＞Ｑ１×ｎ１”が真であった場合（Ｙｅｓ）、遅延障害のディスクドライブと判断し、Ｓ１０７３で閉塞移行処理を実施し、偽であった場合（Ｎｏ）は、正常なディスクドライブと判断し、処理を終了する。
【００９０】

また、前述の比較で、たとえば一例として、“未反映データ量＞Ｑ１×ｎ１÷２”が真であった場合、遅延障害が発生する可能性のあるディスクドライブと判断し、サービスプロセッサインタフェース２５１を介してサービスプロセッサ２５０に通知し、たとえば一例として図１８に示すような画面上において、「ディスクドライブ（Ｄ２）性能遅延発生」のような警告メッセージをサービスプロセッサ２５０上に出力してもよい。

【００９１】

また、本実施の形態のディスクサブシステムでは、通常、ホストコンピュータ４００の一定のサイズ以上のライトデータは均等に分割されることは前述の通りである。均等に分割して格納されたデータは、ホストコンピュータ４００からのリード要求に対し、各ディスクドライブ３１１～３１４より均等に読み出されるということは自明の理である。すなわち、各ディスクドライブ３１１～３１４に対するリード／ライト要求の要求数（キュー数）、転送量はほぼ均一となる。もし、性能的に劣るディスクドライブが存在すれば、そのディスクドライブに対する入出力要求数（キュー数）は他のディスクドライブよりも多くなることになる。

【００９２】

このことより、キュー数を考慮し、このキュー数に着目した応答遅延監視処理を図１９により説明する。図１９では、前記図４に示す構成（３Ｄ＋１Ｐ）でディスクドライブ（Ｄ３）３１３のデータ転送完了後のディスクドライブの遅延処理を例に説明する。

【００９３】

Ｓ１０８１で他のディスクドライブ３１１，３１２，３１４に対するキュー２４１１，２４１２，２４１４の登録数の平均Ｑ２を求める。次に、Ｓ１０８２でディスクドライブ（Ｄ３）３１３のキュー２４１３の登録数とＱ２（Ｑ２×キュー数差分倍数（ｎ２）２２１２）を比較し、“キューの登録数＞Ｑ２×ｎ２”が真であった場合（Ｙｅｓ）、遅延障害のディスクドライブと判断し、Ｓ１０８３で閉塞以降処理を実施し、偽であった場合（Ｎｏ）は、正常なディスクドライブと判断し、処理を終了する。

【００９４】

また、前述の比較で、“キューの登録数＞Ｑ２×ｎ２÷２”が真であった場合、前述の未反映データ量に着目した場合と同様に、遅延障害が発生する可能性のあるディスクドライブと判断し、サービスプロセッサ２５０に通知し、図１８に示すような画面上において、「ディスクドライブ（Ｄ３）性能遅延発生」のような警告メッセージをサービスプロセッサ２５０上に出力してもよい。

【００９５】

また、本実施の形態のディスクサブシステムでは、各ディスクドライブに対する入出力要求数およびデータ転送量は均等であるいうことは前述の通りである。すなわち、ディスクドライブ３１１～３１４の性能が同等であった場合、データ転送にかかる時間（応答時間）も均等になるということは自明の理である。もし、性能的に劣るディスクドライブが存在すれば、そのディスクドライブに対する平均応答時間は他のディスクドライブよりも長くなることになる。

【００９６】

このことより、平均応答時間を考慮し、この平均応答時間に着目した応答遅延監視処理を図２０により説明する。図２０では、前記図４に示す構成（３Ｄ＋１Ｐ）でディスクドライブ（Ｄ１）３１１のデータ転送完了後のディスクドライブの遅延処理を例に説明する。

【００９７】

Ｓ１０９１で他のディスクドライブ３１２，３１３，３１４の共有メモリ２２０上のドライブ情報２２０２，２２０３，２２０４の平均応答時間２３２３の平均Ｑ３を求める。次に、Ｓ１０９２でディスクドライブ（Ｄ１）３１１の平均応答時間２３２３とＱ３（Ｑ３×平均応答時間差分倍数（ｎ３）２２１３）を比較し、“平均応答時間＞Ｑ３×ｎ３”が真であった場合（Ｙｅｓ）、遅延障害のディスクドライブと判断し、Ｓ１０９３で閉塞以降処理を実施し、偽であった場合（Ｎｏ）は、正常なディスクドライブと判断し、処理

を終了する。

【００９８】

また、前述の比較で、“平均応答時間＞Ｑ３×ｎ３÷２”が真であった場合、前述の未反映データ量に着目した場合と同様に、遅延障害が発生する可能性のあるディスクドライブと判断し、サービスプロセッサ２５０に通知し、図１８に示すような画面上において、「ディスクドライブ（Ｄ１）性能遅延発生」のような警告メッセージをサービスプロセッサ２５０上に出力してもよい。

【００９９】

＜ディスクドライブの閉塞移行処理例＞

図２１により、本発明の一実施の形態に係るディスクアレイ装置において、ディスクドライブの閉塞移行処理の一例を説明する。図２１はディスクドライブの閉塞移行処理を示すフローチャートである。図２１では、ディスクドライブ（Ｄ２）３１２の閉塞移行処理を例に説明する。

【０１００】

Ｓ１１０１で閉塞移行するディスクドライブ（Ｄ２）３１２に対応したキュー２４１２を構成する使用情報２４２６のドライブ状態を使用不可能（変更）とし、当該ディスクドライブ３１２に対する入出力要求があった場合は前述の通りとする。このことより、性能劣化のディスクドライブがディスクサブシステムより排除されることになるため、このディスクドライブに対する入出力処理時間の性能劣化が抑制され、結果としてディスクサブシステムとしての性能劣化を回避できる。

【０１０１】

次に、Ｓ１１０２でスペアのディスクドライブ（Ｓ）３１５があるかどうかを判断し、もしスペアのディスクドライブがなかった場合（Ｎｏ）は処理を終了する。Ｓ１１０２でスペアのディスクドライブ３１５がシステムに存在した場合（Ｙｅｓ）は、Ｓ１１０３で他のディスクドライブ３１１，３１３，３１４よりデータリードを行い、キャッシュメモリ２３０上のデータ２３０１，２３０３，２３０４の領域に格納する。次に、Ｓ１１０４でこれらのデータの冗長度よりディスクドライブ（Ｄ２）のデータ２３０５を再生する。続く、Ｓ１１０５で再生したデータ２３０５をスペアのディスクドライブ３１５に書き込む。

【０１０２】

次に、Ｓ１１０６で全てのデータを作成したかどうかを判断し、作成していなかった場合（Ｎｏ）はＳ１１０３に戻る。Ｓ１１０６で全てのデータを作成していた場合（Ｙｅｓ）は、Ｓ１１０７でスペアのディスクドライブ（Ｓ）３１５の使用情報２４２６のドライブ状態を「スペア」から「ディスクドライブ（Ｄ２）として使用可能」に変更し、処理を終了する。

【０１０３】

＜ディスクドライブの性能劣化検出用パラメータの変更処理例＞

図２２～図２５により、本発明の一実施の形態に係るディスクアレイ装置において、ディスクドライブの性能劣化検出用パラメータの変更処理の一例を説明する。それぞれ、図２２は性能遅延の検出レベルを変更するための設定画面を示す図、図２３は検出レベルと各差分係数との対応関係を示す図、図２４は応答時間・入出力処理数を変更するための設定画面を示す図、図２５は検出レベルと実際のキュー数、リトライ回数との対応関係を示す図、である。

【０１０４】

本実施の形態によるディスクサブシステムは、共有メモリ２２０上に格納されるディスクドライブの性能劣化検出用パラメータ２２０６である、未反映データ量の差分倍数（ｎ１）２２１１、キュー数の差分倍数（ｎ２）２２１２、平均応答時間の差分倍数（ｎ３）２２１３の各係数を、サービスプロセッサインタフェース２５１を介して接続されたサービスプロセッサ２５０により変更できる。

【０１０５】

たとえば、図２２に示すようなサービスプロセッサ２５０の設定画面において、ユーザは、性能劣化の検出レベルをＡ（Ｅａｓｙ）、Ｂ（Ｎｏｒｍａｌ）、Ｃ（Ｈａｒｄ）の３段階、および数値を直接入力できるカスタムの計４種類から選択できる。各検出レベルの差分倍数（ｎ１），（ｎ２），（ｎ３）の各係数は、たとえば一例として、図２３に示すような値に設定されている。

【０１０６】

これにより、ユーザは、通常は各係数が中間の値（１．５）のＢの検出レベルを選択し、より早く性能劣化のディスクドライブを検出したい場合は各係数が小さい値（１．２）のＡを、多少の性能劣化があってもディスクドライブを閉塞とさせず、ランニングコストを削減したい場合は各係数が大きい値（２．０）のＣをそれぞれ選択し、より細かい設定を行いたい場合は各係数を任意の値に設定できるカスタムを選択することにより、それぞれのニーズにあったチューニングが実施できる。

【０１０７】

また、本実施の形態によるディスクサブシステムは、ディスクコントローラ２４０のローカルメモリ２４１上に格納されるキュー２４１１～２４１５の数Ｍ１（最大値）をサービスプロセッサ２５０により変更でき、またディスクドライブ３１１～３１５のリトライ回数もサービスプロセッサ２５０により変更できる。

【０１０８】

たとえば、図２４に示すようなサービスプロセッサ２５０の設定画面において、ユーザは、レベルをＡ（応答時間：Ｆａｓｔ、入出力処理数：ＭＩＮ）、Ｂ（応答時間：Ｎｏｒｍａｌ、入出力処理数：Ｎｏｒｍａｌ）、Ｃ（応答時間：Ｓｌｏｗ、入出力処理数：ＭＡＸ）の３段階、および数値を直接入力できるカスタムの計４種類から選択できる。各レベルのキュー数Ｍ１、リトライ回数は、たとえば一例として、図２５に示すような値に設定されている。キュー数を多くした場合は多重性能が上がり、性能遅延障害がない場合には、トランザクション性能が稼げるが、性能遅延障害がある場合には、多重にした分、コマンドが沈み込む可能性がある。

【０１０９】

こうした点より、ユーザは、通常は中間の値（Ｍ１：１、リトライ回数：１０）のＢのレベルを選択し、１コマンドの応答レスポンスを重視し、性能劣化／障害を早急に発見したい場合は小さい値（Ｍ１：１、リトライ回数：５）のＡを、障害がない場合の多重性能を重視したい場合は大きい値（Ｍ１：８、リトライ回数：２０）のＣを選択し、より細かい設定を行いたい場合は任意の値に設定できるカスタムを選択することで、それぞれのニーズにあったチューニングが実施できる。

【０１１０】

従って、本実施の形態のディスクアレイ装置（ディスクサブシステム）によれば、以下のような効果を得ることができる。

【０１１１】

（１）キャッシュメモリ２３０上のディスクドライブに書き込むべきホストコンピュータ４００からの未反映データ量を共有メモリ２２０に保持し、この未反映データ量を冗長度を構成する他のディスクドライブと相対比較することにより、性能劣化のディスクドライブを特定することができる。

【０１１２】

（２）個々のディスクドライブの平均応答時間を共有メモリ２２０に保持し、この平均応答時間を冗長度を構成する他のディスクドライブと相対比較することにより、性能劣化のディスクドライブを特定することができる。

【０１１３】

（３）個々のディスクドライブに対するキュー数をディスクコントローラ２４０に保持し、このキュー数を冗長度を構成する他のディスクドライブと相対比較することにより、性能劣化のディスクドライブを特定することができる。

【０１１４】

（４）性能劣化のディスクドライブを特定し、このディスクドライブを閉塞させた後、冗長性を有する他のディスクドライブのデータを用いてスペアのディスクドライブにデータを復元することができる。

【０１１５】

（５）リトライ回数や、未反映データ量の差分倍数（ｎ１）、キュー数の差分倍数（ｎ２）、平均応答時間の差分倍数（ｎ３）の各係数を変更できるとともに、ユーザが各レベルを選択することができるので、それぞれのニーズにあったチューニングを実施することができる。

【０１１６】

（６）前記（１）～（５）により、ディスクドライブ自身に性能劣化を検出する機能がなくても、性能劣化しているディスクドライブを検出することができ、さらに性能劣化レベルを可変できることにより、ユーザの要求に応じたシステムの構築を実現することができる。

【０１１７】

以上、本発明者によってなされた発明を実施の形態に基づき具体的に説明したが、本発明は前記実施の形態に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることはいうまでもない。

【図面の簡単な説明】

【０１１８】

【図１】本発明の一実施の形態に係るディスクアレイ装置の全体構成を示す構成図である。

【図２】本発明の一実施の形態に係るディスクアレイ装置の外観構成を示す図である。

【図３】本発明の一実施の形態に係るディスクアレイ装置において、ディスク制御装置の外観構成を示す図である。

【図４】本発明の一実施の形態に係るディスクアレイ装置の具体的構成を示す図である。

【図５】本発明の一実施の形態に係るディスクアレイ装置において、ホストコンピュータからのリード要求時のチャネルコントローラの動作を示すフローチャートである。

【図６】本発明の一実施の形態に係るディスクアレイ装置において、チャネルコントローラからのリード要求時のディスクコントローラの動作を示すフローチャートである。

【図７】本発明の一実施の形態に係るディスクアレイ装置において、ホストコンピュータからのライト要求時のチャネルコントローラの動作を示すフローチャートである。

【図８】本発明の一実施の形態に係るディスクアレイ装置において、チャネルコントローラからのライト要求時のディスクコントローラの動作を示すフローチャートである。

【図９】本発明の一実施の形態に係るディスクアレイ装置において、ディスクドライブに対する入出力処理を示す図である。

【図１０】本発明の一実施の形態に係るディスクアレイ装置において、キューの構成を示す図である。

【図１１】本発明の一実施の形態に係るディスクアレイ装置において、ドライブ情報の構成を示す図である。

【図１２】本発明の一実施の形態に係るディスクアレイ装置において、ディスクドライブが使用不可能時のリード要求の動作を示すフローチャートである。

【図１３】本発明の一実施の形態に係るディスクアレイ装置において、ディスクドライブが使用不可能時のライト要求の動作を示すフローチャートである。

【図１４】本発明の一実施の形態に係るディスクアレイ装置において、共有メモリ上

に格納されるディスクドライブの性能劣化検出用パラメータを示す図である。
【図１５】本発明の一実施の形態に係るディスクアレイ装置において、ディスクコントローラとディスクドライブの入出力シーケンスの動作を示すフローチャートである。
【図１６】（ａ）～（ｃ）は本発明の一実施の形態に係るディスクアレイ装置において、具体的な未反映データとデータ転送量の増減の関係を示す図である。
【図１７】本発明の一実施の形態に係るディスクアレイ装置において、未反映データ量に着目した応答遅延監視処理を示すフローチャートである。
【図１８】本発明の一実施の形態に係るディスクアレイ装置において、警告メッセージの表示画面を示す図である。
【図１９】本発明の一実施の形態に係るディスクアレイ装置において、キュー数に着目した応答遅延監視処理を示すフローチャートである。
【図２０】本発明の一実施の形態に係るディスクアレイ装置において、平均応答時間に着目した応答遅延監視処理を示すフローチャートである。
【図２１】本発明の一実施の形態に係るディスクアレイ装置において、ディスクドライブの閉塞移行処理を示すフローチャートである。
【図２２】本発明の一実施の形態に係るディスクアレイ装置において、性能遅延の検出レベルを変更するための設定画面を示す図である。
【図２３】本発明の一実施の形態に係るディスクアレイ装置において、検出レベルと各差分係数との対応関係を示す図である。
【図２４】本発明の一実施の形態に係るディスクアレイ装置において、応答時間・入出力処理数を変更するための設定画面を示す図である。
【図２５】本発明の一実施の形態に係るディスクアレイ装置において、検出レベルと実際のキュー数、リトライ回数との対応関係を示す図である。

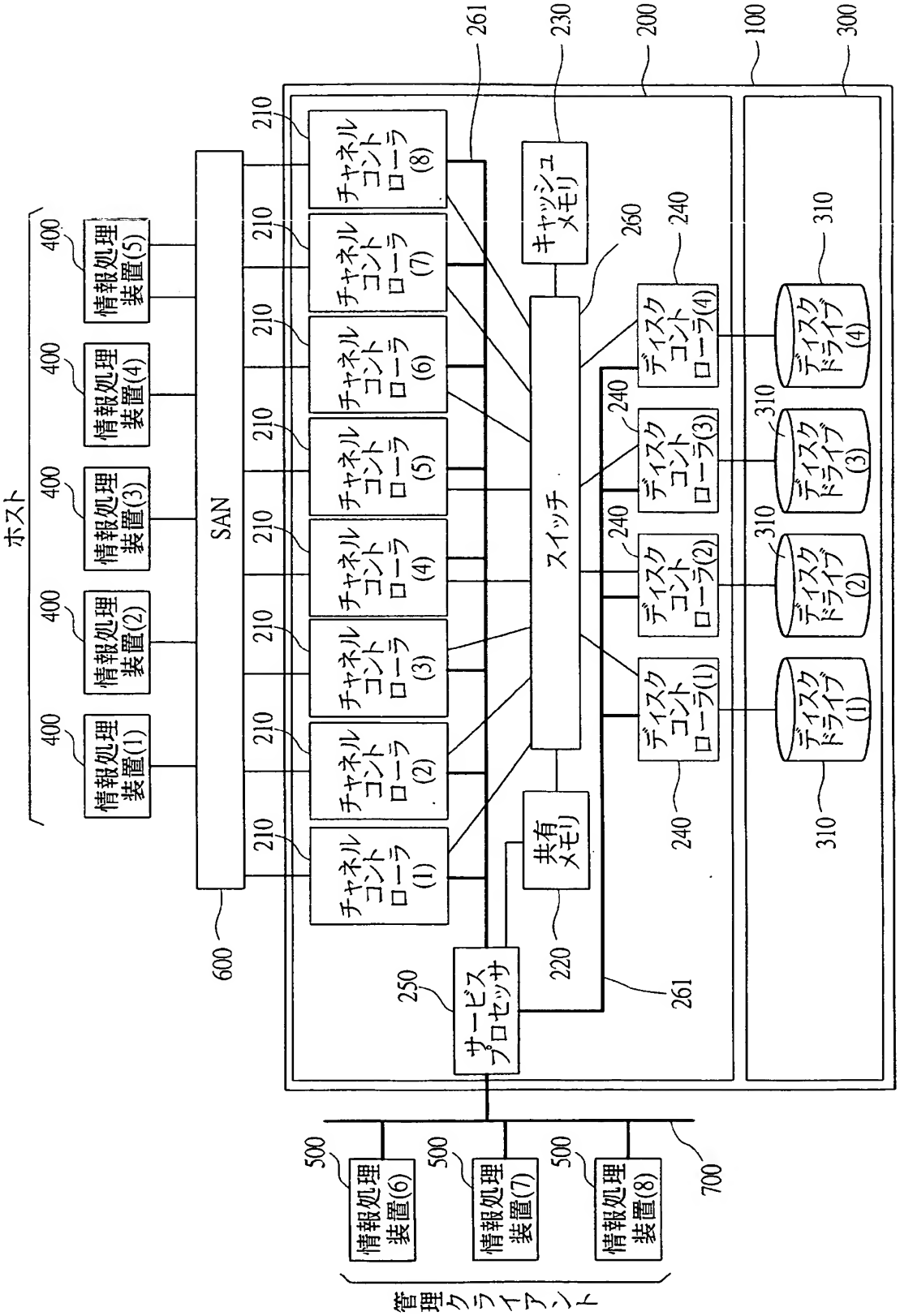
【符号の説明】

【０１１９】

１００…ディスクアレイ装置、２００…ディスク制御装置、２１０…チャネルコントローラ、２２０…共有メモリ、２３０…キャッシュメモリ、２４０…ディスクコントローラ、２４１…ローカルメモリ、２５０…サービスプロセッサ、２５１…サービスプロセッサインタフェース、２６０…スイッチ、２６１…内部ＬＡＮ、２７０…ファン、３００…ディスク駆動装置、３１０，３１１～３１５…ディスクドライブ、４００…情報処理装置（ホストコンピュータ）、５００…情報処理装置（管理クライアント）、６００…ＳＡＮ、７００…ＬＡＮ。

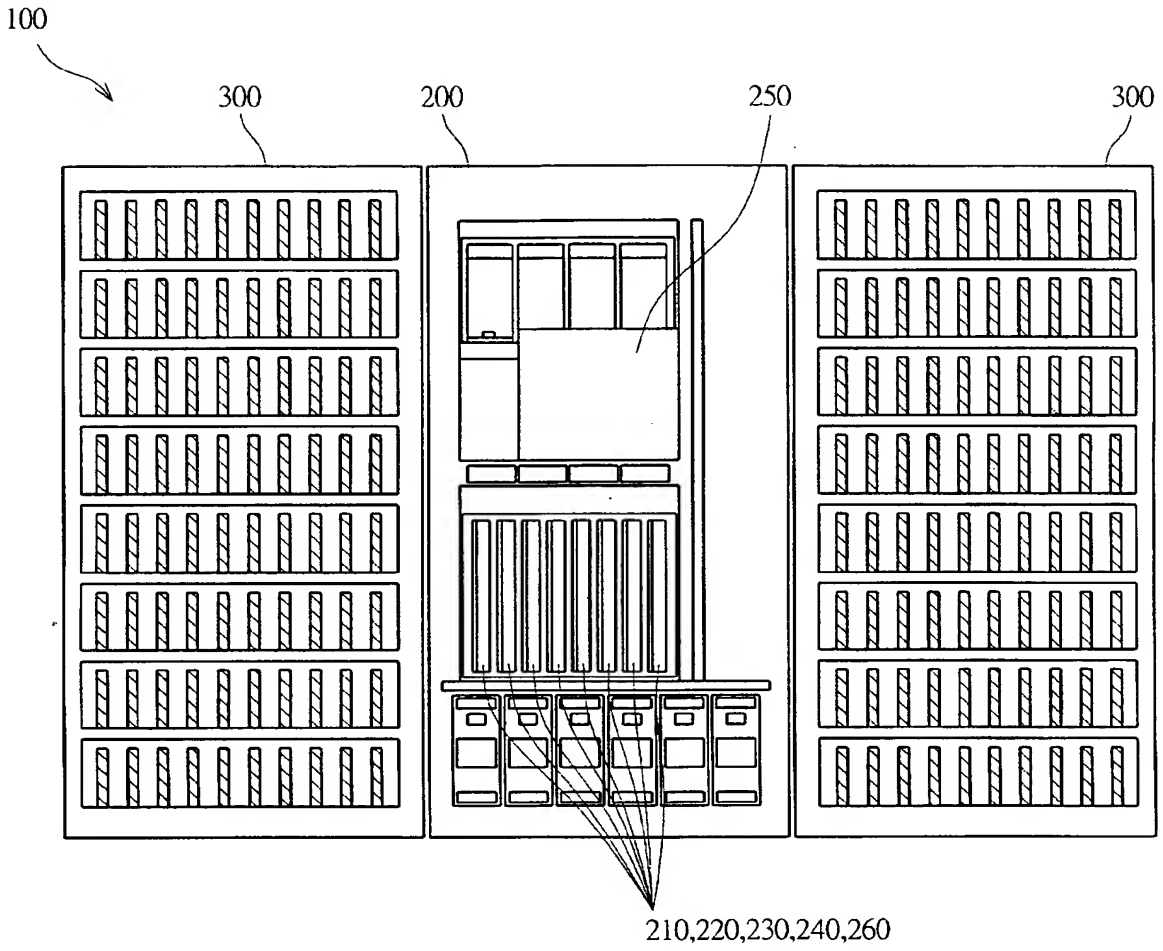
【書類名】図面
【図１】

図１


ホスト
情報処理装置(1)
情報処理装置(2)
情報処理装置(3)
情報処理装置(4)
情報処理装置(5)
SAN
600
チャネルコントローラ(1)
チャネルコントローラ(2)
チャネルコントローラ(3)
チャネルコントローラ(4)
チャネルコントローラ(5)
チャネルコントローラ(6)
チャネルコントローラ(7)
チャネルコントローラ(8)
210
261
230
200
100
300
キャッシュメモリ
スイッチ
260
240
ディスクコントローラ(1)
ディスクコントローラ(2)
ディスクコントローラ(3)
ディスクコントローラ(4)
ディスクドライブ(1)
ディスクドライブ(2)
ディスクドライブ(3)
ディスクドライブ(4)
310
共有メモリ
220
サービスプロセッサ
250
管理クライアント
情報処理装置(6)
情報処理装置(7)
情報処理装置(8)
500
700

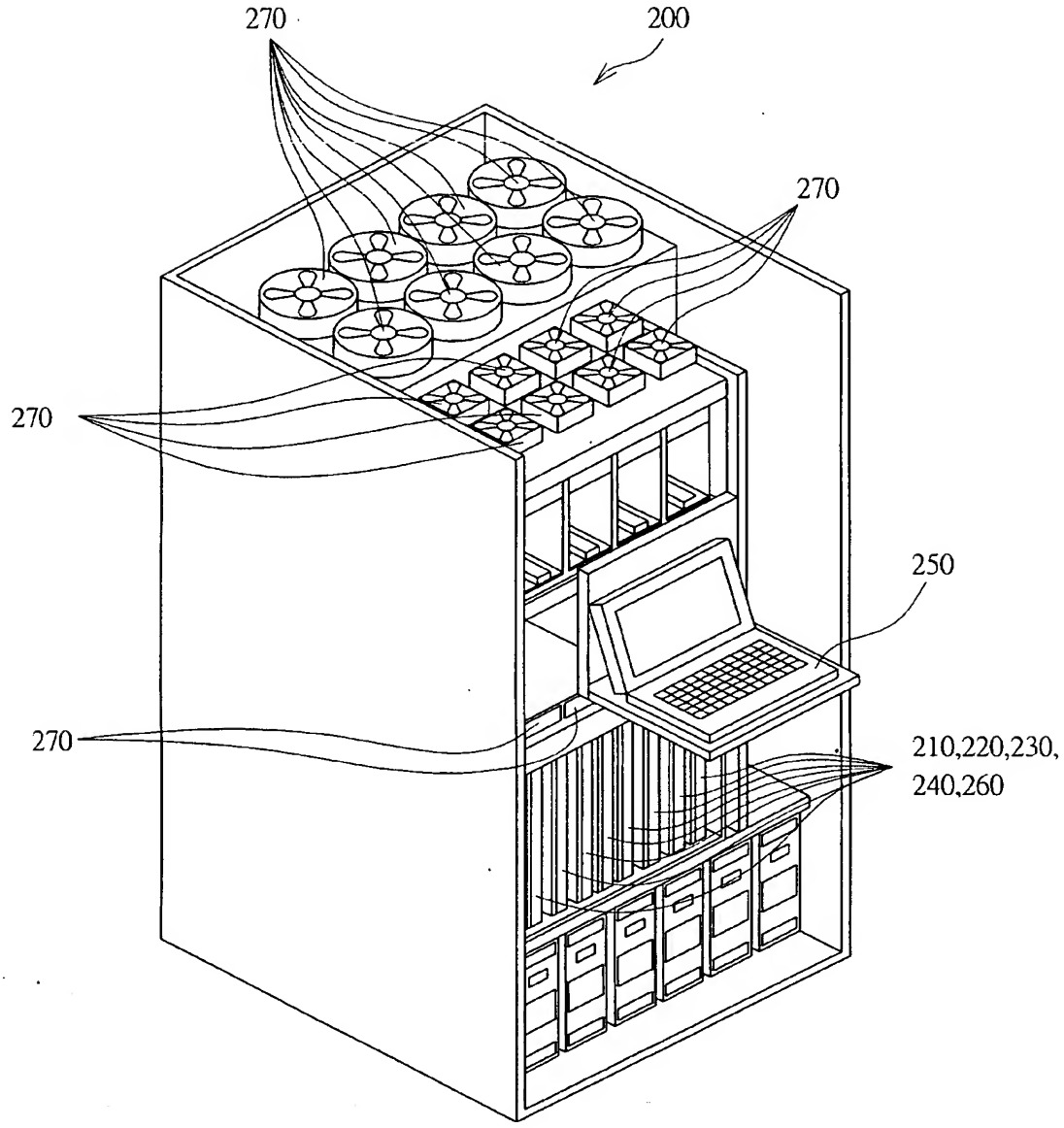
【図２】

図２



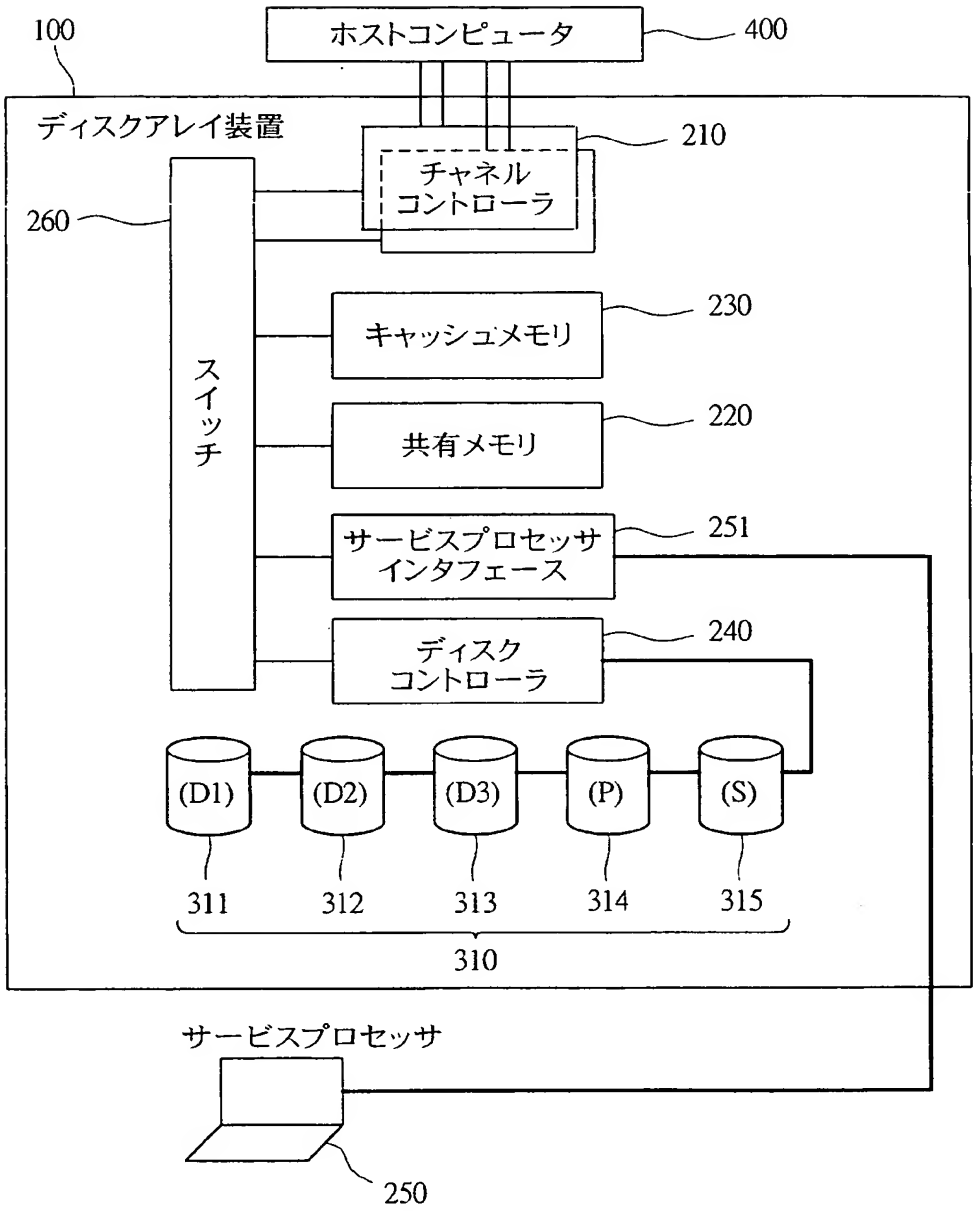
【図３】

図3



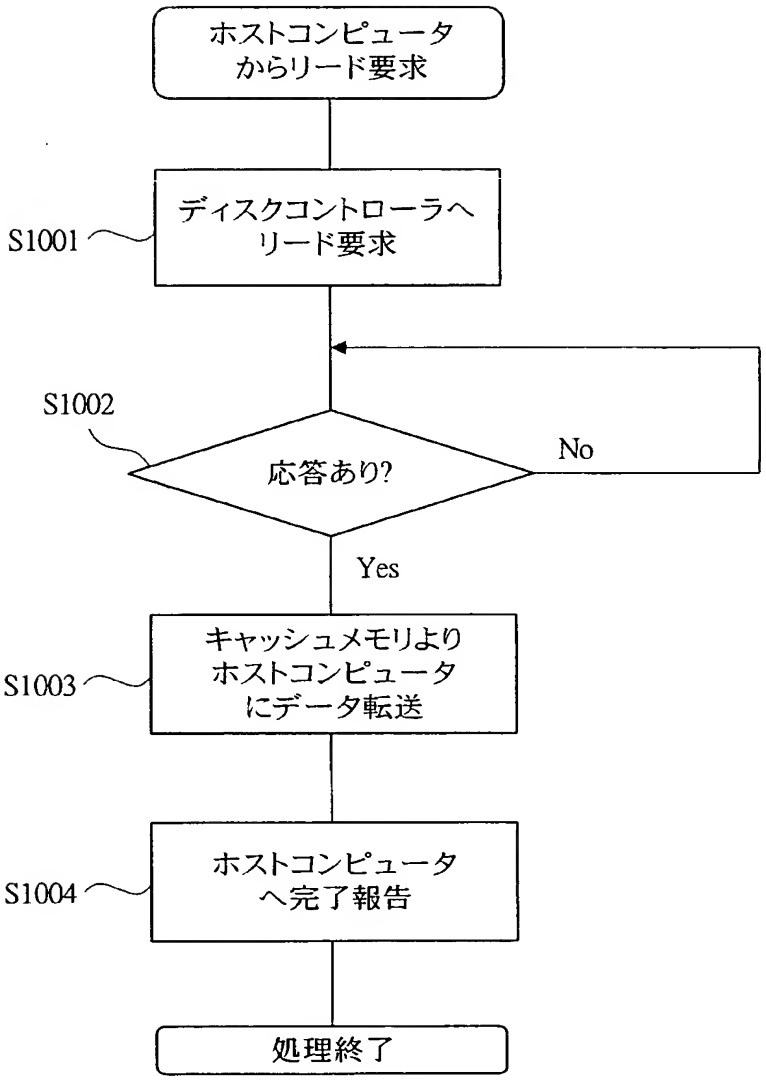
【図４】

図４


100
ホストコンピュータ
400
ディスクアレイ装置
210
チャネル
コントローラ
260
スイッチ
キャッシュメモリ
230
共有メモリ
220
サービスプロセッサ
インタフェース
251
ディスク
コントローラ
240
(D1)
(D2)
(D3)
(P)
(S)
311
312
313
314
315
310
サービスプロセッサ
250

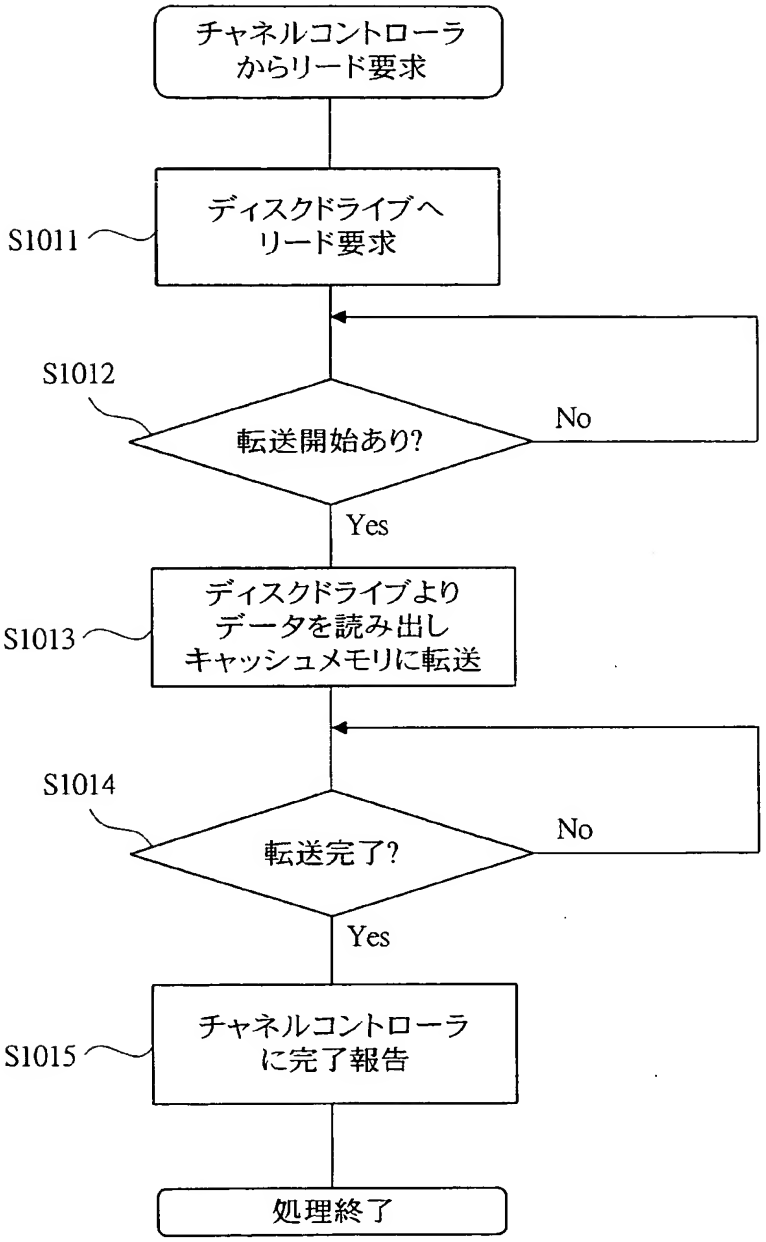
【図５】

図 5


ホストコンピュータ
からリード要求
S1001
ディスクコントローラへ
リード要求
S1002
応答あり?
No
Yes
S1003
キャッシュメモリより
ホストコンピュータ
にデータ転送
S1004
ホストコンピュータ
へ完了報告
処理終了

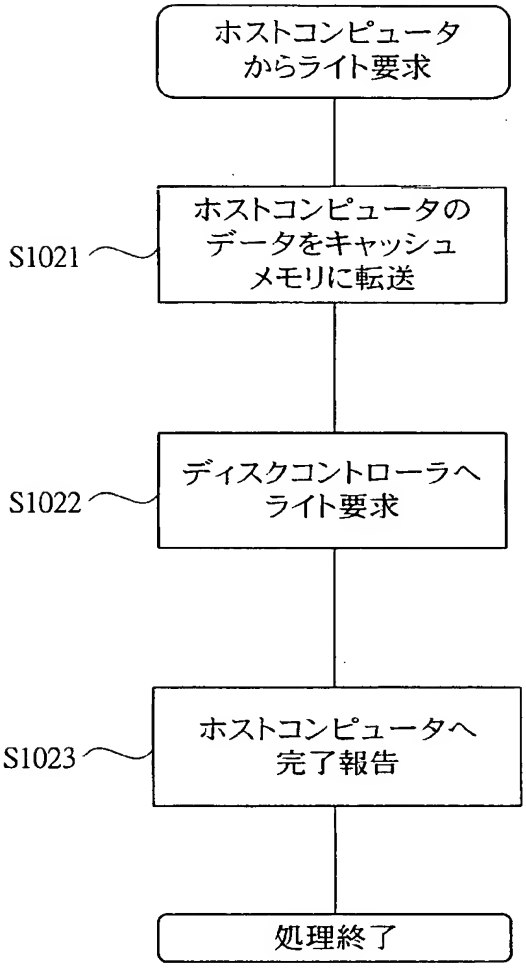
【図６】

図 6


チャネルコントローラからリード要求
S1011
ディスクドライブへリード要求
S1012
転送開始あり?
No
Yes
S1013
ディスクドライブよりデータを読み出しキャッシュメモリに転送
S1014
転送完了?
No
Yes
S1015
チャネルコントローラに完了報告
処理終了

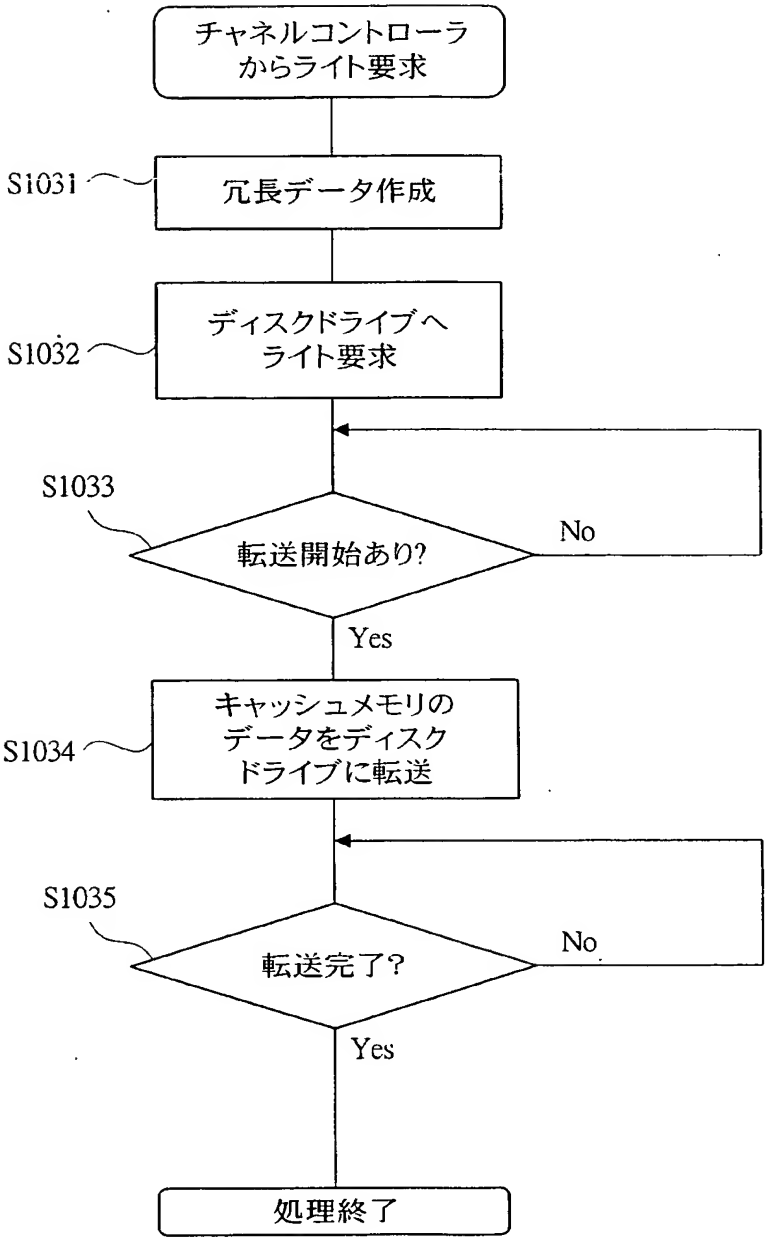
【図７】

図 7


ホストコンピュータ
からライト要求
S1021
ホストコンピュータの
データをキャッシュ
メモリに転送
S1022
ディスクコントローラへ
ライト要求
S1023
ホストコンピュータへ
完了報告
処理終了

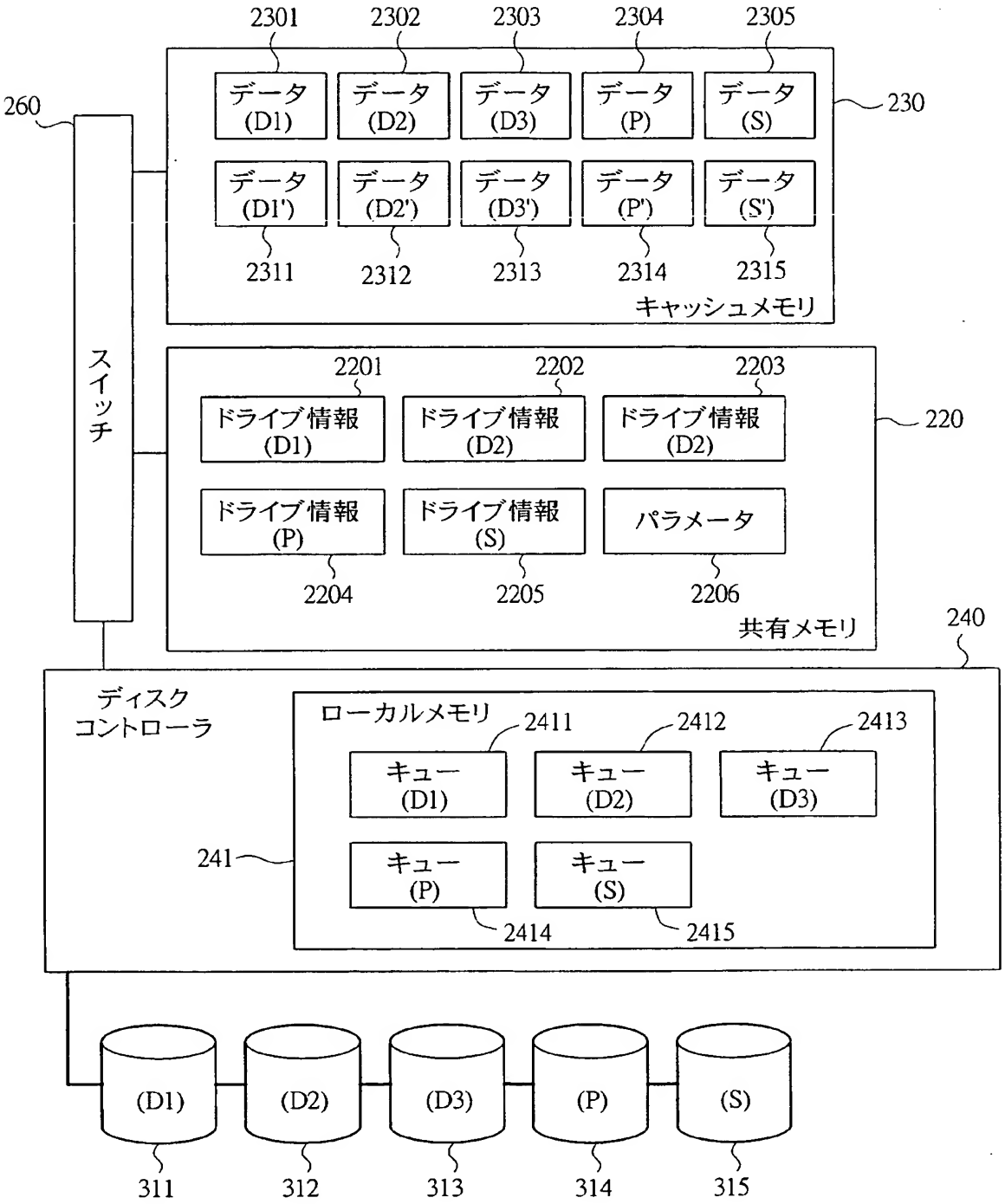
【図８】

図 8


チャネルコントローラからライト要求
S1031
冗長データ作成
S1032
ディスクドライブへライト要求
S1033
転送開始あり?
No
Yes
S1034
キャッシュメモリのデータをディスクドライブに転送
S1035
転送完了?
No
Yes
処理終了

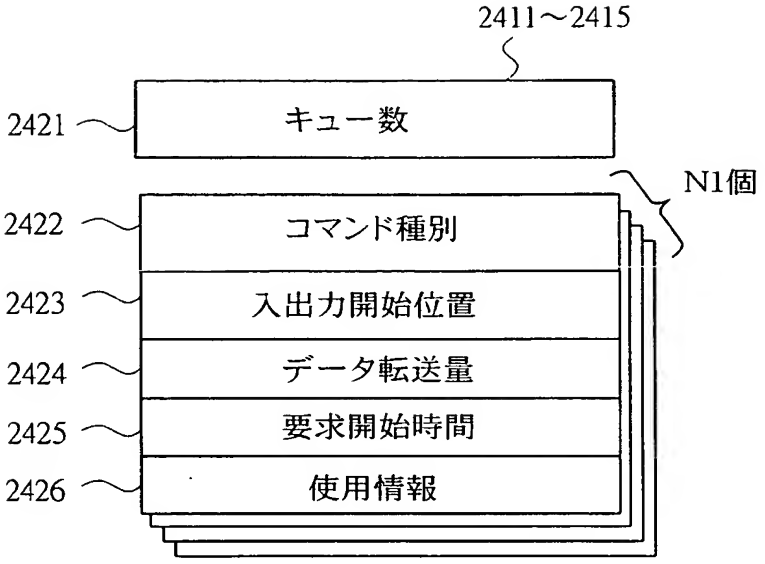
【図９】

図９


2301
2302
2303
2304
2305
データ (D1)
データ (D2)
データ (D3)
データ (P)
データ (S)
データ (D1')
データ (D2')
データ (D3')
データ (P')
データ (S')
2311
2312
2313
2314
2315
キャッシュメモリ
230
260
スイッチ
2201
2202
2203
ドライブ情報 (D1)
ドライブ情報 (D2)
ドライブ情報 (D2)
ドライブ情報 (P)
ドライブ情報 (S)
パラメータ
2204
2205
2206
共有メモリ
220
240
ディスク コントローラ
ローカルメモリ
2411
2412
2413
キュー (D1)
キュー (D2)
キュー (D3)
241
キュー (P)
キュー (S)
2414
2415
(D1)
(D2)
(D3)
(P)
(S)
311
312
313
314
315

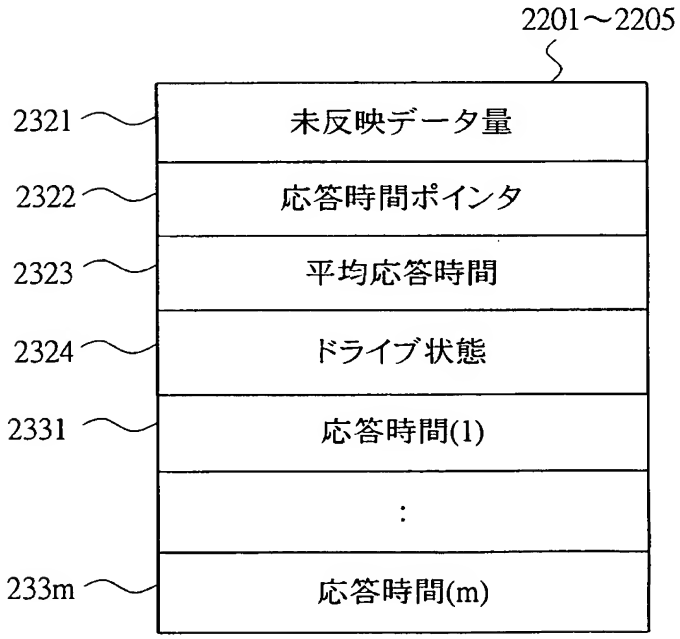
【図１０】

図 10



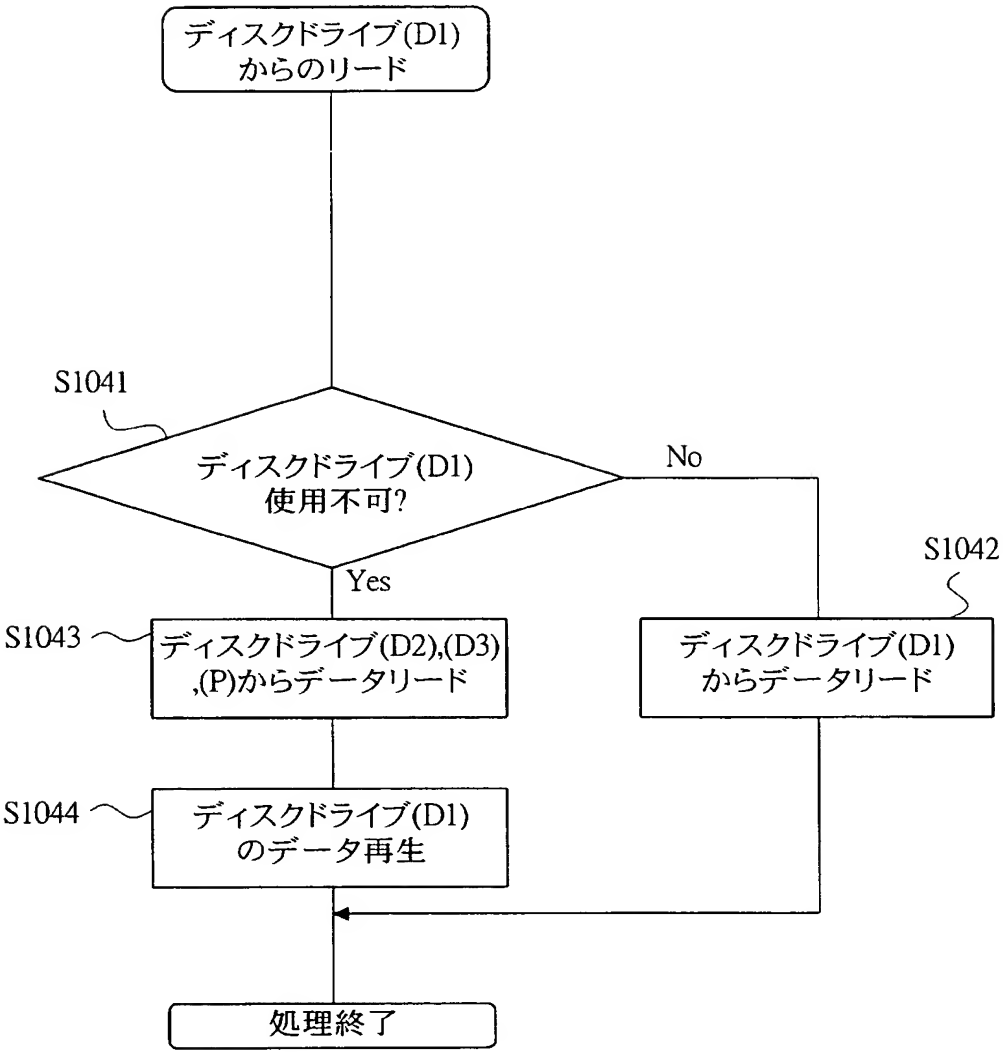
【図１１】

図 11



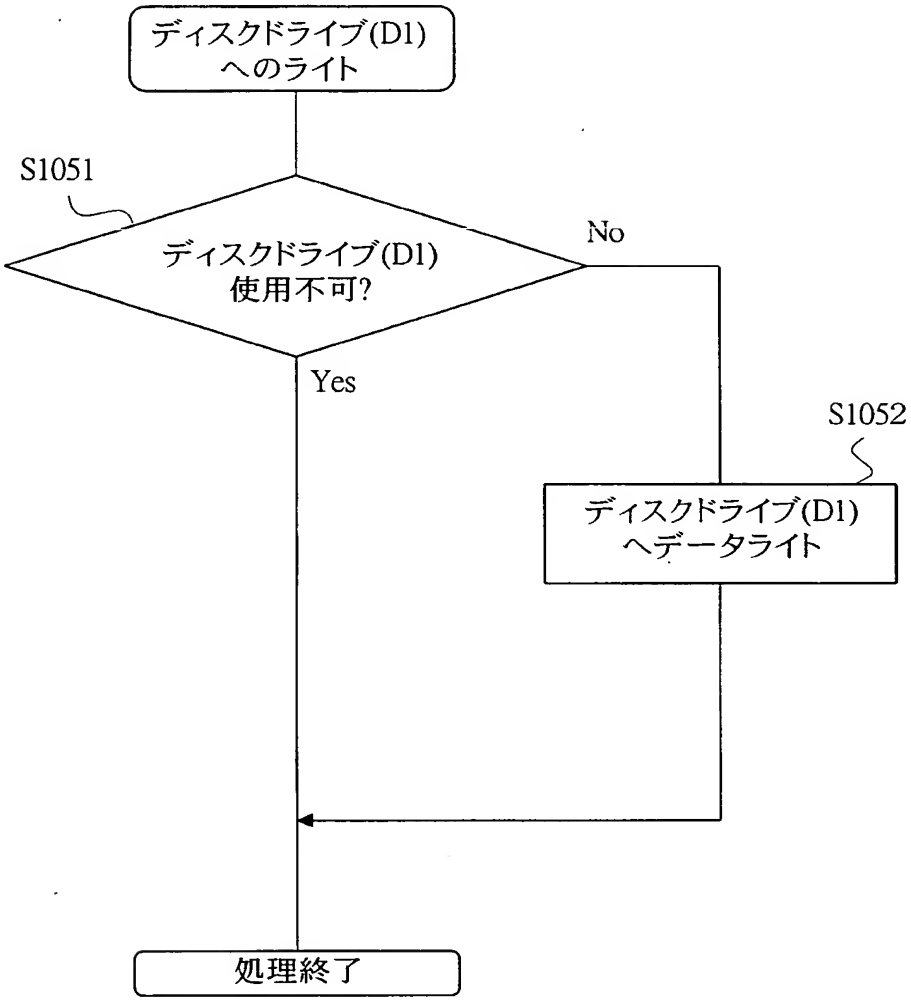
【図１２】

図 12


ディスクドライブ(D1)
からのリード
S1041
ディスクドライブ(D1)
使用不可?
No
Yes
S1042
S1043
ディスクドライブ(D2),(D3)
,(P)からデータリード
ディスクドライブ(D1)
からデータリード
S1044
ディスクドライブ(D1)
のデータ再生
処理終了

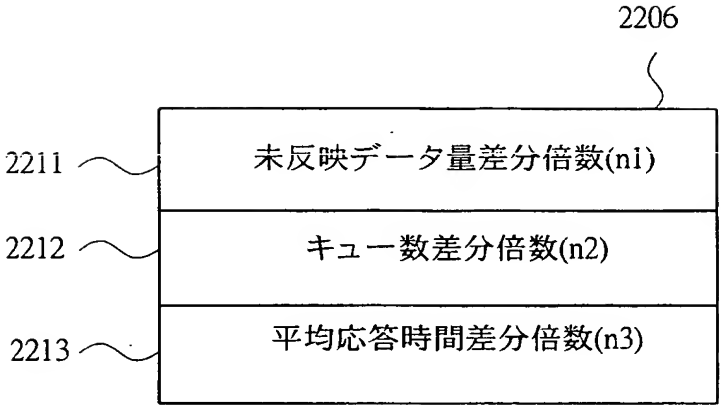
【図１３】

図 13


ディスクドライブ(D1)
へのライト
S1051
ディスクドライブ(D1)
使用不可?
No
Yes
S1052
ディスクドライブ(D1)
へデータライト
処理終了

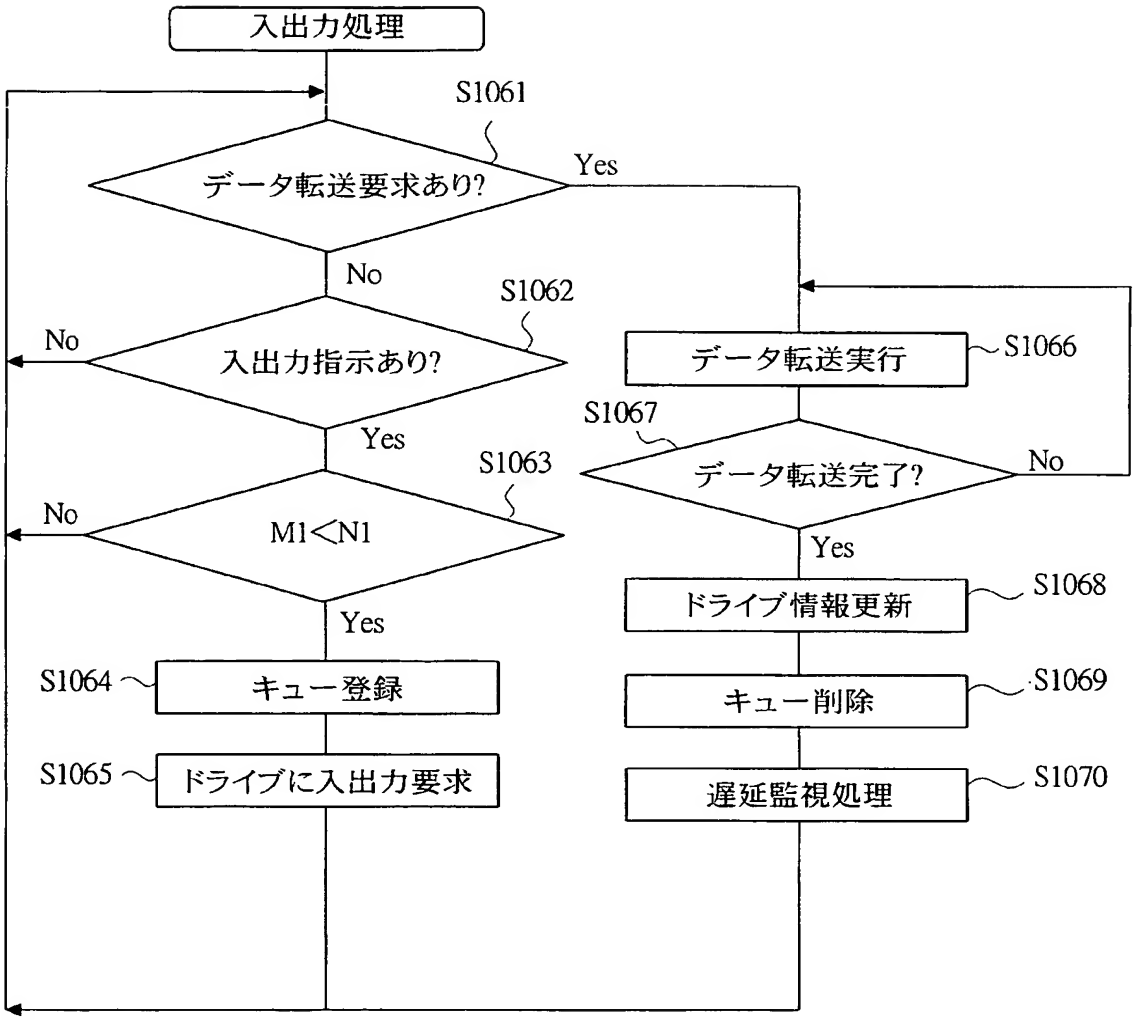
【図１４】

図 14


2206
2211
未反映データ量差分倍数(n1)
2212
キュー数差分倍数(n2)
2213
平均応答時間差分倍数(n3)

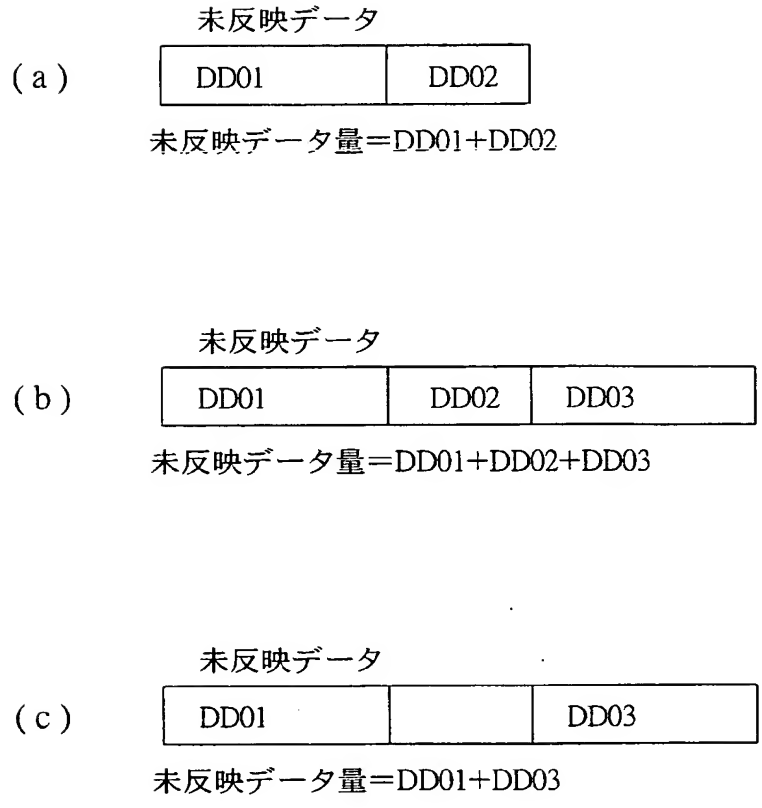
【図１５】

図 15


入出力処理
S1061
データ転送要求あり?
Yes
No
S1062
No
入出力指示あり?
Yes
S1063
No
M1<N1
Yes
S1064
キュー登録
S1065
ドライブに入出力要求
S1066
データ転送実行
S1067
データ転送完了?
No
Yes
S1068
ドライブ情報更新
S1069
キュー削除
S1070
遅延監視処理

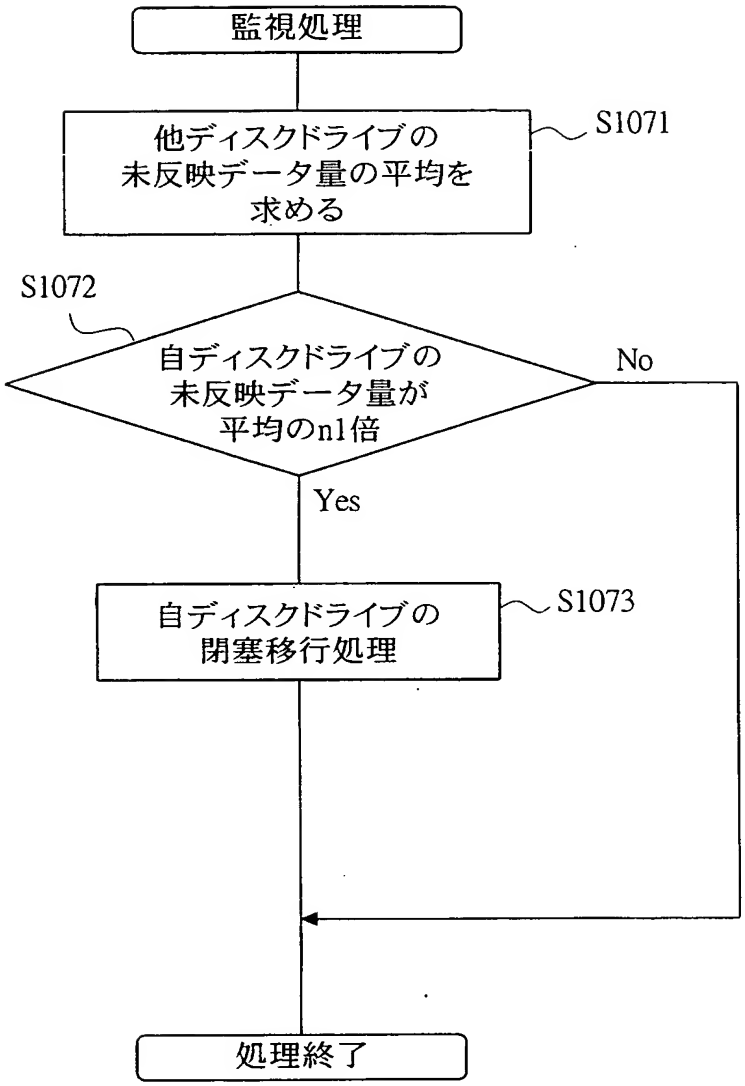
【図１６】

図 16


未反映データ
(a)
DD01
DD02
未反映データ量＝DD01+DD02
未反映データ
(b)
DD01
DD02
DD03
未反映データ量＝DD01+DD02+DD03
未反映データ
(c)
DD01
DD03
未反映データ量＝DD01+DD03

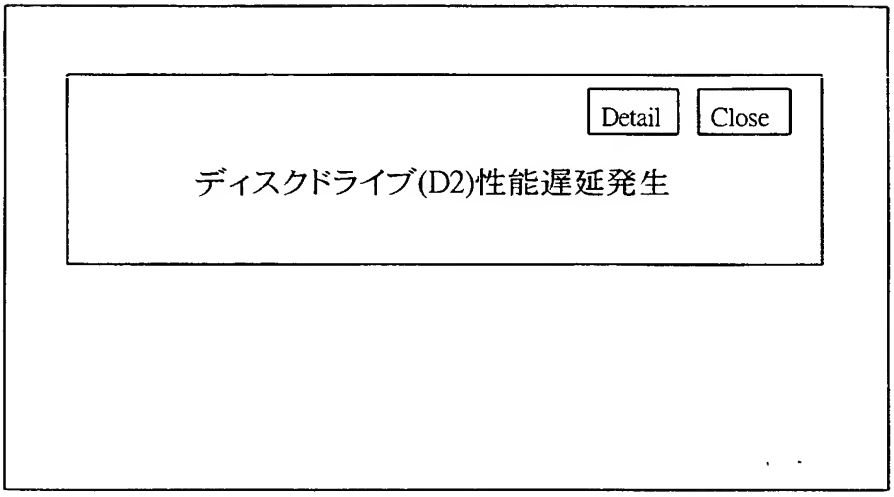
【図１７】

図 17


監視処理
他ディスクドライブの
未反映データ量の平均を
求める
S1071
S1072
自ディスクドライブの
未反映データ量が
平均のn1倍
No
Yes
自ディスクドライブの
閉塞移行処理
S1073
処理終了

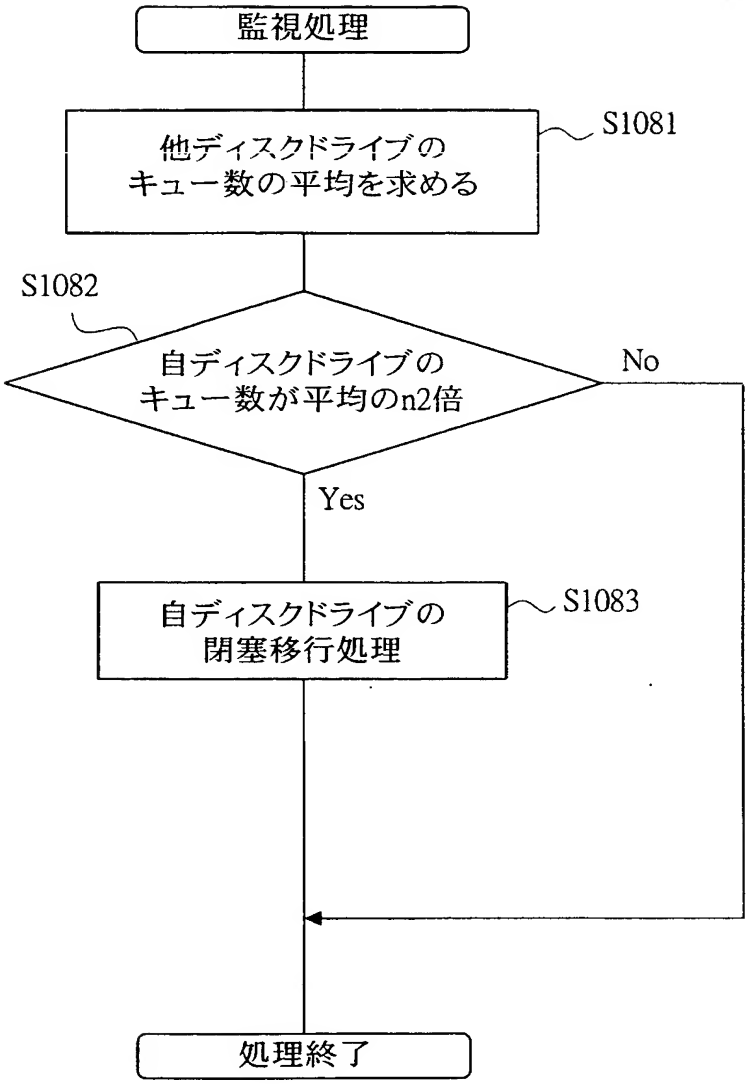
【図１８】

図 18



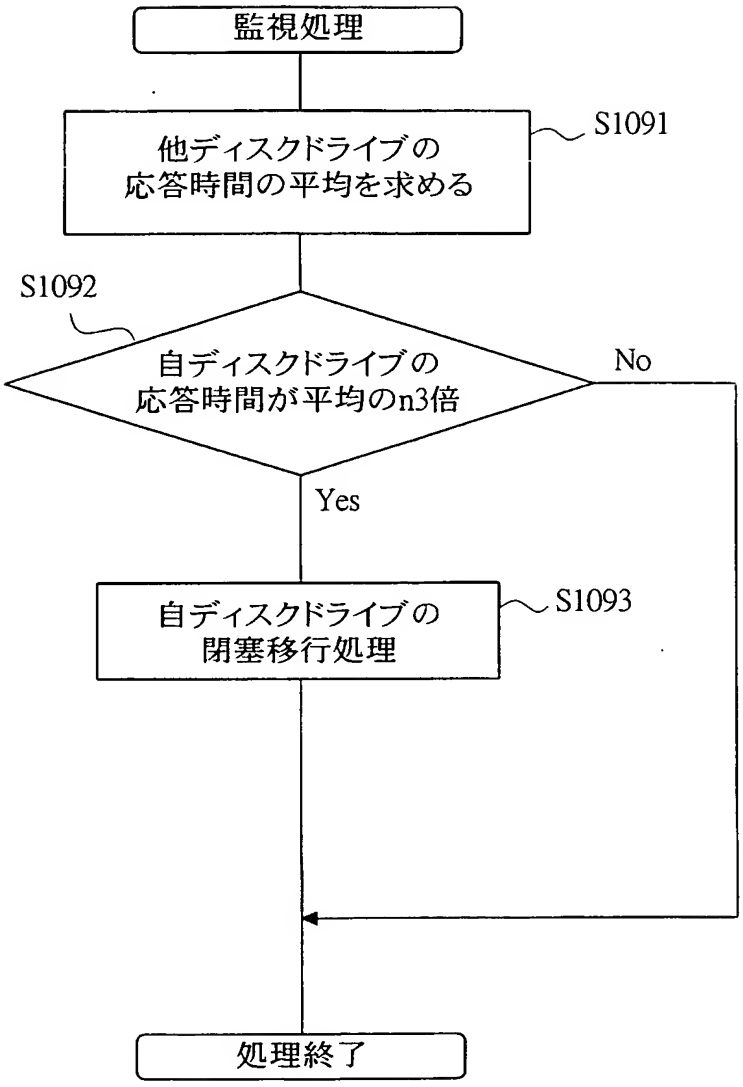
【図１９】

図 19


監視処理
他ディスクドライブの
キュー数の平均を求める
S1081
S1082
自ディスクドライブの
キュー数が平均のn2倍
No
Yes
自ディスクドライブの
閉塞移行処理
S1083
処理終了

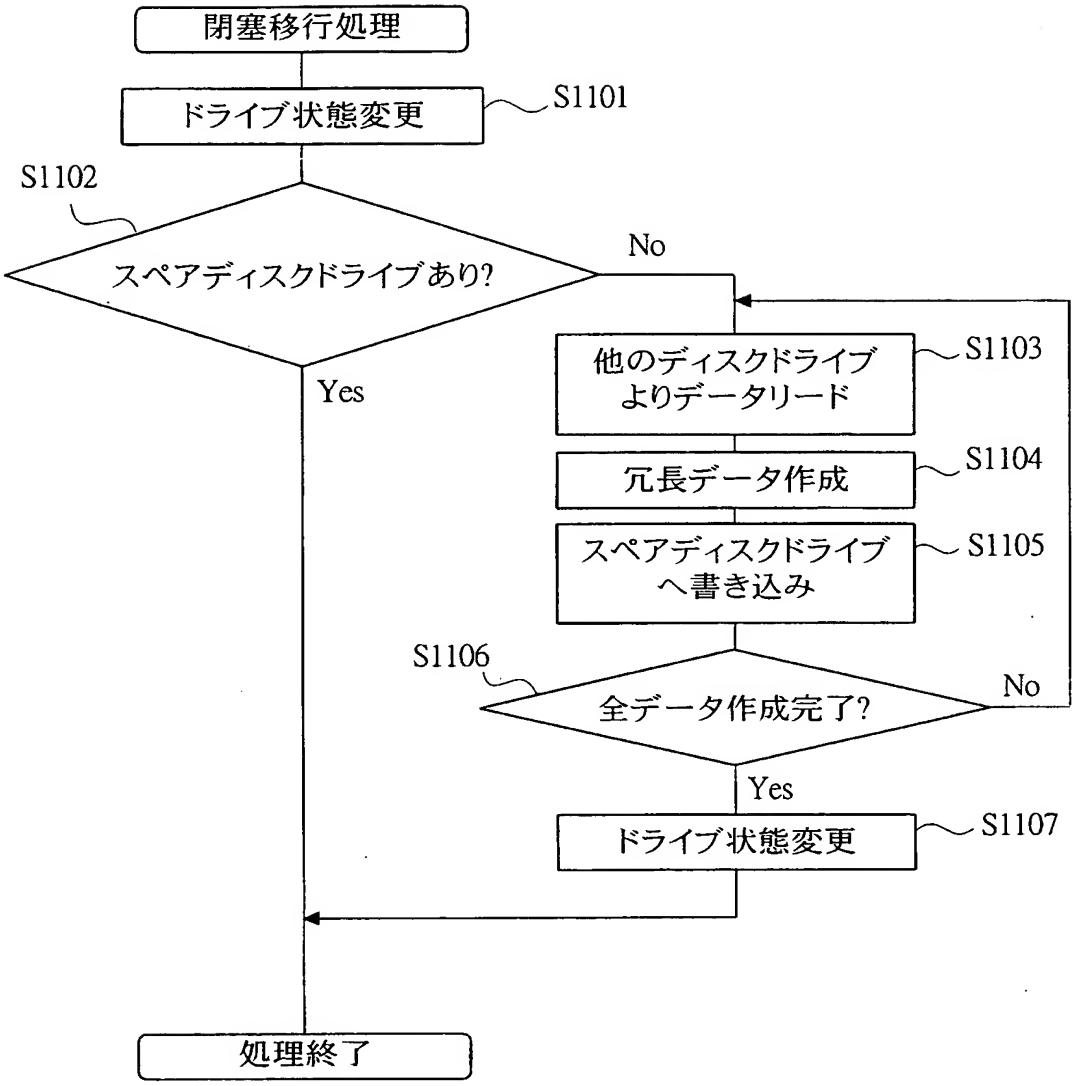
【図２０】

図 20


監視処理
他ディスクドライブの
応答時間の平均を求める
S1091
S1092
自ディスクドライブの
応答時間が平均のn3倍
No
Yes
自ディスクドライブの
閉塞移行処理
S1093
処理終了

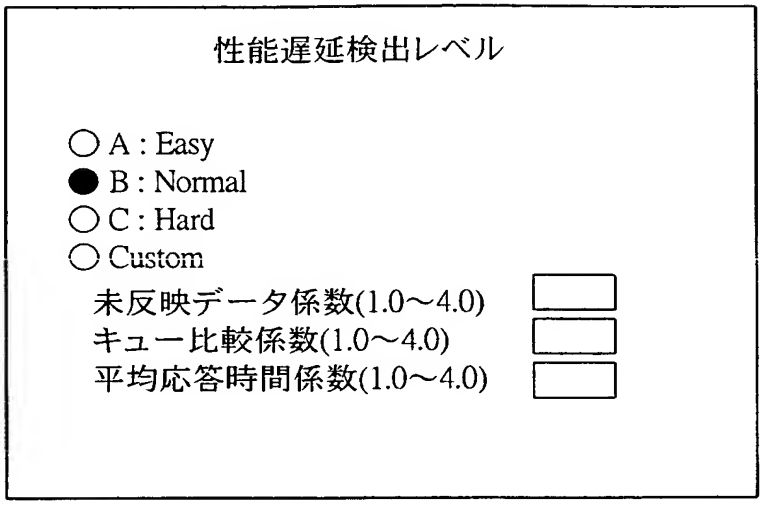
【図２１】

図 21


閉塞移行処理
ドライブ状態変更
S1101
S1102
スペアディスクドライブあり?
No
Yes
他のディスクドライブ
よりデータリード
S1103
冗長データ作成
S1104
スペアディスクドライブ
へ書き込み
S1105
S1106
全データ作成完了?
No
Yes
ドライブ状態変更
S1107
処理終了

【図２２】

図 22



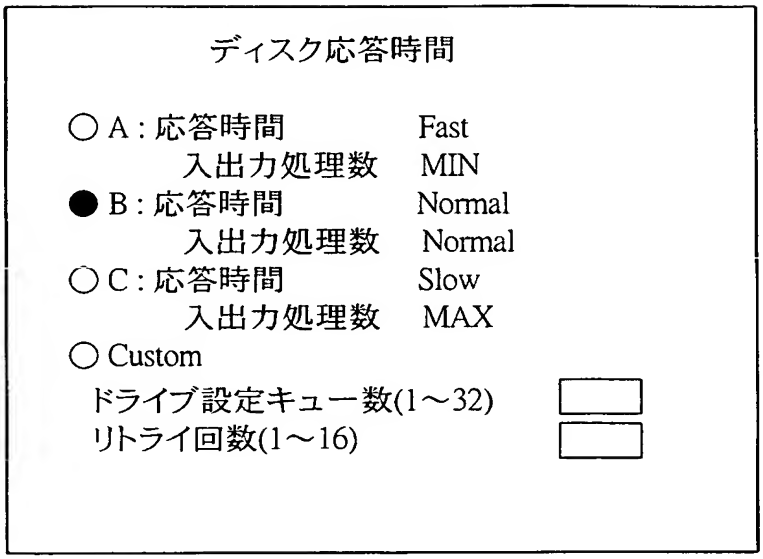
【図２３】

図 23

レベル	n1	n2	n3
A	1.2	1.2	1.2
B	1.5	1.5	1.5
C	2.0	2.0	2.0
カスタム	任意	任意	任意

【図２４】

図 24



【図２５】

図 25

レベル	M1	リトライ回数
A	1	5
B	4	10
C	8	20
カスタム	任意	任意

【書類名】要約書
【要約】
【課題】　ディスクドライブ自身に性能劣化を検出する機能がなくても、性能劣化しているディスクドライブを検出することができ、さらに性能劣化レベルを可変できることにより、顧客要求に応じたディスクアレイ装置の構築を実現できる技術を提供する。
【解決手段】　ディスクアレイ装置１００において、ディスクコントローラ２４０は、データの書き込みまたは読み出しに利用され、冗長性を有してデータを格納できる論理的な記憶領域を、複数のディスクドライブの記憶領域を用いて生成し、論理的な記憶領域を形成する複数のディスクドライブに関して、複数のディスクドライブに対する書き込みデータまたは読み出し要求が格納される格納領域を監視し、論理的な記憶領域を形成する複数のディスクドライブのうちデータの書き込みまたは読み出しの繰り返し回数が多いディスクドライブを特定して、特定されたディスクドライブを閉塞させる。
【選択図】　図４

特願２００４－０６３３１３

出　願　人　履　歴　情　報

識別番号　［０００００５１０８］

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所